Office of Technology Assessment

Congressional Board of the 100th Congress

MORRIS K. UDALL, *Arizona, Chairman*

TED STEVENS, *Alaska, Vice Chairman*

Senate

ORRIN G. HATCH
Utah

CHARLES E. GRASSLEY
Iowa

EDWARD M. KENNEDY
Massachusetts

ERNEST F. HOLLINGS
South Carolina

CLAIBORNE PELL
Rhode Island

House

GEORGE E. BROWN, JR.
California

JOHN D. DINGELL
Michigan

CLARENCE E. MILLER
Ohio

DON SUNDQUIST
Tennessee

AMO HOUGHTON
New York

JOHN H. GIBBONS
(Nonvoting)

Advisory Council

WILLIAM J. PERRY, *Chairman*
H&Q Technology Partners

DAVID S. POTTER, *Vice Chairman*
General Motors Corp. (Ret.)

EARL BEISTLINE
Consultant

CHARLES A. BOWSHER
General Accounting Office

S. DAVID FREEMAN
Lower Colorado River Authority

MICHEL T. HALBOUTY
Michel T. Halbouty Energy Co.

NEIL E. HARL
Iowa State University

JAMES C. HUNT
University of Tennessee

JOSHUA LEDERBERG
Rockefeller University

CHASE N. PETERSON
University of Utah

SALLY RIDE
Stanford University

JOSEPH E. ROSS
Congressional Research Service

Director

JOHN H. GIBBONS

The Technology Assessment Board approves the release of this report. The views expressed in this report are not necessarily those of the Board, OTA Advisory Council, or individual members thereof.

COVER DESIGN BY JOHN BERGLING

MAPPING
OUR GENES

The Genome Projects: How Big, How Fast?

CONGRESS OF THE UNITED STATES   OFFICE OF TECHNOLOGY ASSESSMENT

WASHINGTON, DC 20510-8025

J952-40

Recommended Citation:

U.S. Congress, Office of Technology Assessment, *Mapping Our Genes—The Genome Projects: How Big, How Fast?* OTA-BA-373 (Washington, DC: U.S. Government Printing Office, April 1988).

Library of Congress Catalog Card Number 87-619 898

For sale by the Superintendent of Documents
U.S. Government Printing Office, Washington, DC 20402-9325
(order form can be found in the back of this report)

Foreword

For the past 2 years, scientific and technical journals in biology and medicine have extensively covered a debate about whether and how to determine the function and order of human genes on human chromosomes and when to determine the sequence of molecular building blocks that comprise DNA in those chromosomes. In 1987, these issues rose to become part of the public agenda. The debate involves science, technology, and politics. Congress is responsible for "writing the rules" of what various Federal agencies do and for funding their work. This report surveys the points made so far in the debate, focusing on those that most directly influence the policy options facing the U.S. Congress.

The House Committee on Energy and Commerce requested that OTA undertake the project. The House Committee on Science, Space, and Technology, the Senate Committee on Labor and Human Resources, and the Senate Committee on Energy and Natural Resources also asked OTA to address specific points of concern to them. Congressional interest focused on several issues:

- how to assess the rationales for conducting human genome projects,
- how to fund human genome projects (at what level and through which mechanisms),
- how to coordinate the scientific and technical programs of the several Federal agencies and private interests already supporting various genome projects, and
- how to strike a balance regarding the impact of genome projects on international scientific cooperation and international economic competition in biotechnology.

OTA prepared this report with the assistance of several hundred experts throughout the world. Their help included interviews with OTA staff, comments on drafts of the report, and sending information to OTA. We want to thank those reviewers and many others who have contributed to making the report more accurate, balanced, and useful.

This report is one of many OTA reports related to biotechnology and genetics. Recent reports on related topics are *Technologies for Detecting Heritable Mutations in Human Beings*, *New Developments in Biotechnology: 1) Ownership of Human Tissues and Cells, 2) Public Perceptions of Biotechnology, 4) U.S. Investment in Biotechnology*, and *Human Gene Therapy*.

John H Gibbons

JOHN H. GIBBONS
Director

OTA Advisory Panel on Mapping the Human Genome

LeRoy Walters, Ph.D., *Chairman*
Director, Center for Bioethics
Kennedy Institute of Ethics, Georgetown University

George F. Cahill, M.D.
Vice President for Scientific Training and Development
Howard Hughes Medical Institute

Susan E. Cozzens, Ph.D.
Assistant Professor
Department of Science and Technology Studies
Rensselaer Polytechnic Institute of Technology

Tamara J. Erickson, Ph.D.
Vice President
Health Industries
Arthur D. Little, Inc.

Joseph G. Gall, Ph.D.
Department of Embryology
Carnegie Institution of Washington

Walter B. Goad, Ph.D.
Theoretical Biology Group
Los Alamos National Laboratory

Leroy Hood, Ph.D., M.D.
Chairman
Division of Biology
California Institute of Technology

Horace Freeland Judson
Henry R. Luce Professor
The Writing Seminars
The Johns Hopkins University

William W. Lowrance, Ph.D.
Director, Life Sciences and Public Policy Program
The Rockefeller University

Norman R. Pace, Ph.D.
Professor of Biology
Department of Biology
Indiana University

Mark L. Pearson, Ph.D.
Director of Molecular Biology
E.I. du Pont de Nemours & Co.

George Rose, Ph.D.
Professor
Department of Biological Chemistry
Hershey Medical Center
Pennsylvania State University

Margery W. Shaw, M.D., J.D.
Professor of Health Law
University of Texas Health Science Center

Dieter Soll, Ph.D.
Professor
Department of Molecular Biophysics and Biochemistry
Yale University

Nancy S. Wexler, Ph.D.
President
Hereditary Disease Foundation and
Associate Professor of Clinical Neuropsychology
Departments of Neurology and Psychiatry
Columbia University

Raymond L. White, Ph.D.
Investigator
Howard Hughes Medical Institute, and Professor
Department of Human Genetics
University of Utah School of Medicine

NOTE: OTA appreciates and is grateful for the valuable assistance and thoughtful critiques provided by the advisory panel members. The panel does not, however, necessarily approve, disapprove, or endorse this report. OTA assumes full responsibility for the report and the accuracy of its contents.

OTA Human Genome Project Staff

Roger Herdman, *Assistant Director, OTA*
Health and Life Sciences Division

Gretchen Kolsrud, Biological Applications Program Manager

Robert Mullan Cook-Deegan, *Project Director*
Patricia Hoben, *Analyst*
Jacqueline Courteau, *Research Assistant*
Gladys White, *Analyst*
David Guston, *OTA Summer Intern*

OTA Support Staff

Sharon K. Oatman, *Administrative Assistant*
Linda S. Rayford, *Secretary/Word Processing Specialist*
Barbara V. Ketchum, *Clerical Assistant*

Editor

Blair Burns Potter

Graphics

MedSciArtCo, Washington, DC

Contractors

(for list of topics and availability of reports see app. A)

Computer Horizons, Inc.
Theodore Friedmann, University of California, San Diego
Jonathan Glover, New College, Oxford University
Allen Hammond, Consultant, Washington, DC
John Heilbron, University of California, Berkeley, and
Daniel Kevles, California Institute of Technology
Horace Freeland Judson, The Johns Hopkins University
Marc Lappé, University of Illinois at Chicago
Steve Mount, Yale University
Teresa Myers, Consultant, Washington, DC
Richard Myers, University of California, San Francisco
Peter Newmark, London
Susan Rosenfeld, Science and the Law Committee, Association of the Bar of the City of New York
David Weatherall, Oxford University
Akihiro Yoshikawa, Berkeley Roundtable on the International Economy, University of California, Berkeley

OTA Staff Reviewers

L. Val Giddings   Kathi Hanna   Lisa Heinz
Kevin O'Connor   Gary Ellis   Mark Schaefer

Contents

Page

Chapter 1: Summary 3

Chapter 2: Technologies for Mapping DNA 21

Chapter 3: Applications to Research in Biology and Medicine 55

Chapter 4: Social and Ethical Considerations 79

Chapter 5: Agencies and Organizations in the United States 93

Chapter 6: Organization of Projects 115

Chapter 7: International Efforts 133

Chapter 8: Technology Transfer 165

Appendix A: Topics of OTA Contract Reports 179

Appendix B: Participants in OTA Workshops 180

Appendix C: Estimated Costs of Human Genome Projects 187

Appendix D: Databases, Repositories, and Informatics 189

Appendix E: Bibliometric Analysis of Human Genome Research 195

Appendix F: Acknowledgments 196

Appendix G: Glossary 201

Index 207

Chapter 1

Summary

CONTENTS

Page
Debates About Mapping the Human Genome 4
The Focus of Genome Projects 7
Misplaced Controversy About "The Human Genome Project" 9
The Core Issue: Resource Allocation for Research Infrastructure 10
Organization of This Report 11
The Role of Congress 11
Options for Action by Congress 11
  Appropriations to Federal Agencies 11
  Access to Information and Materials 12
  Organization of Genome Projects 12
  Technology Transfer 15
  Questions for Congressional Oversight 17

Figure

Figure *Page*
1-1. Comparative Scale of Mapping 5

Table

Table *Page*
1-1. Principal Organizations Involved in Genome Projects 7

Chapter 1

Summary

> "We want the maximum good per person; but what is good? To one person it is wilderness, to another it is ski lodges for thousands. To one it is estuaries to nourish ducks for hunters to shoot at; to another it is factory land. Comparing one good with another is, we usually say, impossible because goods are incommensurable. Incommensurables cannot be compared.
>
> Theoretically this may be true; but in real life, *incommensurables are commensurable*. All that is needed is a criterion of judgment and a system of weighing."
>
> Garret Hardin, "The Tragedy of the Commons,"
> *Science* 162:1243-1248, 1968.

> "Congress is the place where we make impossible choices between apples and oranges. We do it every year in preparing the largest budget on the planet."
>
> Congressional staff member, 1988.

> "All legislative powers granted shall be vested in a Congress of the United States . . . . No money shall be drawn from the Treasury, but in consequence of appropriations made by law. . . ."
>
> Article 1, U.S. Constitution.

The mysteries of inheritance are surrendering to modern biology. Over a century ago, Austrian monk Gregor Mendel demonstrated that the inheritance of traits could be most simply explained if it were controlled by factors passed from one generation to the next. These units of inheritance came to be called genes. The complete set of genes from an organism is called its genome. Some traits are best explained by inheritance of single genes (e.g., many genetic diseases, colorblindness), but most, including many nongenetic diseases, involve combinations of multiple genes with environmental factors.

Scientists discovered in the 1940s that genes consisted of DNA (deoxyribonucleic acid), and in the 1950s they further elucidated the mechanisms of inheritance. In 1953, Watson and Crick described the structure of DNA—the double helix—which provides at once an explanation of how genetic material is inherited and how genes direct cellular function. DNA encodes the blueprint for every living thing; it is packed into chromosomes which can be seen under a light microscope. The genome of an organism can thus be defined as the DNA comprising its chromosomes. Each human cell has 46 chromosomes in 23 pairs. One chromosome of each pair is inherited from each parent. DNA consists of long chains of chemicals called nucleotide bases. There are four such bases, represented most simply as A, C, T, and G. The order of bases making up DNA is called its sequence. The DNA sequence contains the instructions that specify the production of molecules, usually proteins, that provide cellular structure and perform biochemical functions in the cell.

Our understanding of genetics has advanced remarkably in the last three decades as new methods of manipulating and analyzing DNA have been developed. Recombinant DNA technology enables scientists to insert DNA from one organism directly into that of another, thereby allowing them to study how genes function in relatively controlled conditions. New methods to detect and purify small amounts of DNA, new techniques to handle and analyze DNA that is millions of bases long, and novel scientific instruments have augmented the tools scientists use to understand heredity. These powerful and rapidly evolving technologies have provoked debate in recent years about whether and how to mount a concerted research program to map the human genome and to determine its DNA sequence.

To date, the combined efforts of government agencies, university researchers, and private supporters of biomedical research have produced rough but extremely useful maps of DNA markers covering most regions of the human chromosomes. Chromosomal locations of over 1,215 human genes are now known (of the 50,000 to 150,000 estimated to exist), including those causing all 20 of the most common genetic diseases. Sequencing of DNA from human beings has increased sharply in recent years, yet far fewer than 1 percent of the more than 3 billion bases comprising the human genome have been sequenced (see figure 1-1). The function of only a few hundred human genes is known. Some genetic disorders are understood at the molecular level (e.g., sickle cell disease and Tay-Sachs disease), but the mechanisms underlying most genetic diseases remain unknown. Genetic factors underlying other diseases are known only in barest outline.

The growing power and speed of research in molecular biology have led to proposals to apply novel molecular biological methods to the genetics of entire organisms. **Research and technology efforts aimed at mapping and sequencing large portions or entire genomes are called genome projects.** These proposals would build on experience already gained from mapping lower organisms (e.g., yeast, nematodes, and bacteria) and sequencing some virus genomes and regions of other organisms, yet they would be more ambitious in scale and complexity. More specifically, a public debate began in 1985 about the feasibility of mapping, and perhaps sequencing, the human genome and that of certain other organisms. The debate has often been cast as an on-off decision about whether there should be a concerted Federal effort, yet this is an oversimplification. There are many component projects at different stages of completion: Systematically making maps of human chromosomes is a continuation of ongoing efforts, for example. Databases for genetic information and repositories for research materials are essential whether or not there are other special efforts. Developing new technologies is widely agreed to be important and will require focused research programs. The most contentious issue is whether the DNA sequence of all human chromosomes should be determined. There is little doubt that large regions of human chromosomes will be sequenced eventually, but there is vigorous debate about whether a massive, concerted sequencing effort is warranted. This remains an open question that is likely to be resolved only after pilot projects to determine the sequence of other organisms, small human chromosomes, or chromosomal regions of special interest have been performed. Pilot projects can demonstrate the technologies and should also determine whether dedicated sequencing efforts are efficient and scientifically sensible.

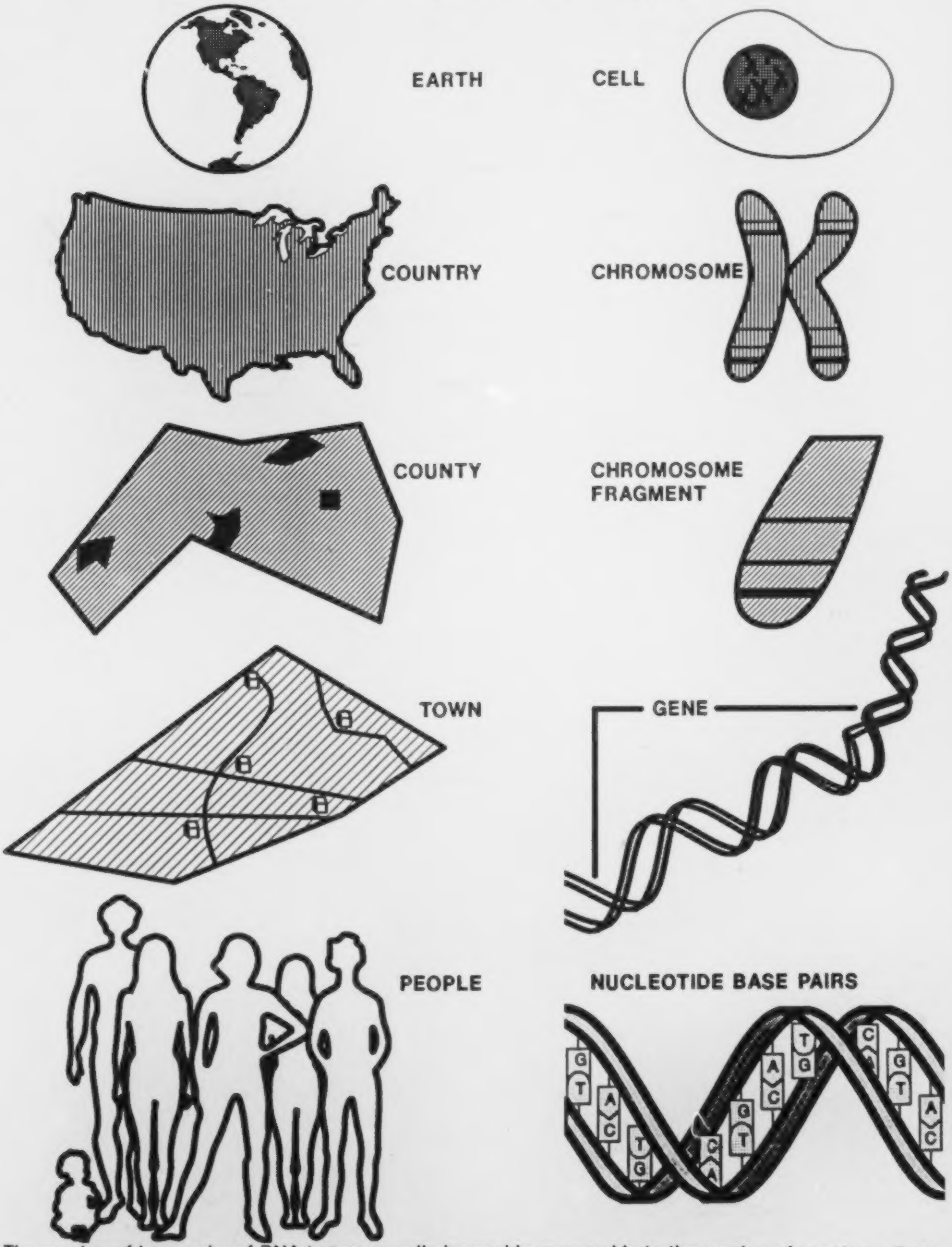
Two scientific advisory groups—one reporting to the Department of Energy (DOE) and the other convened by the National Research Council (NRC) of the National Academy of Sciences—recommended augmented funding of $200 million per year for genome projects. An Office of Technology Assessment (OTA) workshop attempted to estimate the costs of major component projects. Projections fell into the range of $45 to $50 million per year initially, increasing to $200 to $250 million per year over 5 years. Funding recommendations made by the scientific advisory committees would cover most but not all costs estimated by OTA.

DEBATES ABOUT MAPPING THE HUMAN GENOME

The debate about mapping the human genome can be traced through several phases. Until the 1960s, techniques for locating human genes were rudimentary, and human genetics was based primarily on analysis of inheritance patterns of diseases and other observable traits through family trees. In the late 1960s and through the 1970s, scientists developed the first maps of human genes, based on direct observation of chromosomes. In successful cases, the location of a gene could be specified within several million bases of DNA.

In the late 1970s and early 1980s, scientists took the first steps toward maps of human chromosomes based on direct biochemical analysis of

Figure 1-1.—Comparative Scale of Mapping


EARTH
CELL
COUNTRY
CHROMOSOME
COUNTY
CHROMOSOME FRAGMENT
TOWN
GENE
PEOPLE
NUCLEOTIDE BASE PAIRS

The number of base pairs of DNA in human cells is roughly comparable to the number of people on Earth. The scale of genetic mapping efforts can be compared to population maps, with chromosomes (50 to 250 million base pairs) analogous to nations, and genes (thousands to millions of base pairs) to towns.

DNA. DNA fragments of unknown function but known location were used to study inheritance of traits far more precisely than before. Calculations suggested that DNA markers, which signify the presence or absence of particular stretches of DNA, could be identified for regions of all the human chromosomes.

Markers can be used to trace which pieces of DNA, and therefore which parts of chromosomes, are inherited from which parent. When a genetic trait is caused by a single gene and that gene is close to a marker, the marker can be used to ascertain roughly where the gene is located because the two are inherited together.

The U.S. Government and research agencies abroad fund most research that uses DNA markers to study diseases and physiological functions and most university groups searching for new markers in chromosomal regions of particular interest. In the United States, the National Institutes of Health (NIH) are the largest funding sources for biomedical research on genetics.

Construction of maps of DNA markers was undertaken in the early 1980s. The two largest collections of markers were developed by the Howard Hughes Medical Institute (HHMI), a private philanthropy, and Collaborative Research, Inc., a private corporation. Dozens of university researchers and other private firms also contributed to this kind of genetic map.

In 1985, DOE began planning the Human Genome Initiative to develop research tools for molecular genetics. Events leading up to the initiative included a workshop convened by the University of California at Santa Cruz and internal planning by DOE administrators. DOE considered the initiative an extension of its ongoing work in molecular biology—largely focused on detecting mutations and other biological effects of radiation and energy production—that would take advantage of research staff and instruments located at the national laboratories, which are funded by DOE. DOE held several public meetings to discuss the technical possibilities. The first of these was a workshop held in March 1986 in Santa Fe, New Mexico.

Discussion at that workshop of whether to establish a reference sequence for the entire human genome touched off a controversy that has persisted ever since. Arguments about the usefulness of extensive sequence information reached a high pitch at a conference at Cold Spring Harbor Laboratories in June 1986. Many scientists perceived a major sequencing effort as a threat to the conduct of basic research in molecular biology because of its projected cost and potential drain on research talent. Estimates of the cost of sequencing alone (without accounting for mapping or preparation of DNA to be sequenced) ran to billions of dollars. Calls for central management of such a prodigious undertaking further heightened tension because of the strong tradition of decentralized, small-group research in molecular biology. Debate over the appropriate strategy for deciding which regions to sequence first added to the din and spilled over into the scientific press. Major newspapers and magazines have covered the debate since, giving the Human Genome Initiative a high public profile.

The Cold Spring Harbor discussion was followed by a series of meetings held by HHMI, NIH, DOE, NRC, OTA, and others. Plans for special research initiatives by NIH, DOE, and HHMI have resulted from these and other discussions. A few private corporations have also been established (or are being established) to perform DNA sequencing and to develop research resources.

This report deals with various projects that have been proposed by Federal agencies to construct maps of human and other chromosomes, to improve relevant databases and repositories, and to improve research methods and instruments. **There is no single human genome project, but instead many projects**. For 1988, there are specific line items in appropriations for DOE and NIH, and the bulk of the discussion in this report refers to these new research programs. For purposes of this report, ***genome projects*** **refers to the research programs of NIH, DOE, and HHMI, as well as parallel programs in the private sector or other nations**.

THE FOCUS OF GENOME PROJECTS

Genome projects have several objectives:

- to establish, maintain, and enhance databases containing information about DNA sequences, location of DNA markers and genes, function of identified genes, and other related information;
- to create maps of human chromosomes consisting of DNA markers that would permit scientists to locate genes quickly;
- to create repositories of research materials, including ordered sets of DNA fragments that fully represent DNA in the human chromosomes;
- to develop new instruments for analyzing DNA;
- to develop new ways to analyze DNA, including biochemical and physical techniques and computational methods;
- to develop similar resources for other organisms that would facilitate biomedical research; and possibly
- to determine the DNA sequence of a large fraction of the human genome and that of other organisms.

Genome projects underway or planned by DOE, NIH, the National Science Foundation (NSF), HHMI, and other organizations are different but overlapping. They share two features: They would put new methods and instruments into the tool kit of molecular biology, and they would build a research infrastructure for geneticists (see table 1-1).

DOE's Human Genome Initiative began in late 1986 and consists of several projects. One is to create an ordered set of DNA segments from known chromosomal locations; this set, if widely available, could save the tedious steps involved in isolating DNA for study once a gene's approximate location is known. It should also reduce needless duplication of effort by different groups studying genes in the same chromosomal region. A second project is to develop new computational methods to enhance analysis of genetic map and DNA sequence data. Another project is to develop new techniques and instruments for detecting and analyzing DNA, including automation and robotics. For these projects, DOE expended $4.2 million in 1987 and plans $12 million for 1988. It also planned to support an additional $7 million in 1987

Table 1-1.—Principal Organizations Involved in Genome Projects

Organization	Mission	Funding ($000,000s)[a]
National Institutes of Health (Department of Health and Human Services)	Biomedical research	Life sciences: 6,170 Related research: 313 Genome projects: 17.2 NLM biotechnology databases: 3.83
Department of Energy (Office of Health and Environmental Research, Office of Energy Research)	Biological effects of energy production and radiation; use of national laboratory resources	Life sciences: 230 Related research: 7 Genome projects: 12
National Science Foundation (Directorate of Biological, Behavioral, and Social Sciences	Basic scientific research	Life sciences: 206 Related research: 32.7 Genome projects: 0.2
Howard Hughes Medical Institute	Biomedical research	Life sciences: 240 Genetics: 40 Genetic marker maps: 2 to 4 Databases: 2

[a] "Life sciences" figures are estimates for fiscal year 1987; these are total budgets for NIH and HHMI and estimates of relevant programs for NSF and DOE. Figures for "related research" include basic research projects that involve mapping or sequencing, and research infrastructure such as databases and repositories. "Related research" figures are estimates for fiscal year 1987. "Genome projects" figures are estimates for fiscal year 1988, based on appropriations under the December 1987 continuing resolution.

SOURCES: NIH: Rachel Levinson, personal communications, October, November, December 1987, January 1988; DOE: David Smith, personal communications, June, October 1987, January 1988; NSF: David Kingsbury, personal communications, June and November 1987; HHMI: George Cahill, personal communication, January 1988.

for related research and infrastructure. DOE has requested $18.5 million for direct support of its Human Genome Initiative in fiscal year 1989.

NIH has supported special genome projects since 1987, with two objectives: to improve methods for analyzing the genome of human beings and other complex organisms and to enhance computational methods. NIH also supports most of the relevant databases and repositories. It spent an estimated $313 million on projects that involved mapping and sequencing in 1987, and several million more on infrastructure. NIH plans somewhat higher spending for related research in 1988 and will have two items in its budget—an additional $17.2 million for genome projects and $3.83 million for increased database support at the National Library of Medicine. The fiscal year 1989 budget request for the National Institute of General Medical Sciences of NIH includes $28 million for genome projects.

HHMI has two genome initiatives: one to support key databases containing information about the genetics of human and other organisms, and the other to support biomedical research on basic genetic mechanisms and genetic disease. HHMI's budget estimates from 1987 included $40 million for genetics (including $2 to $4 million for genetic mapping) and $2 million to support genome databases.

The NSF plans to increase the number of biology centers it supports, in order to develop new scientific instrumentation and encourage sharing of expensive equipment. These and other NSF programs are not genome projects per se, but they are likely to be integrated with programs of other agencies in some locations. Instrumentation developed through the biology centers will probably be directly relevant to genome projects. NSF budget estimates for 1987 were $206 million for life sciences, of which $32.7 million went to research related to genome projects and $200,000 went directly to genome projects.

Mechanisms for interagency coordination of genome projects have evolved over the past 2 years. Initially, there was informal communication among DOE, NIH, NSF, and HHMI. The Federal agencies then formed a working group under the Domestic Policy Council (DPC), a cabinet-level group in the White House. A committee to replace the DPC group is now being organized by the White House Office of Science and Technology Policy (OSTP), but its exact composition and function have not yet been determined.

International efforts are concentrated in developed nations with strong research traditions. Mapping genes, both human and nonhuman, has been an international effort since its inception. International agreements for databases (particularly those containing DNA sequence data) and collaborations on gene mapping (notably, the Center for the Study of Human Polymorphism in Paris) have been in operation for several years. **No foreign government has made a commitment yet to mapping and sequencing the human genome,** although several governments support related projects through their usual mechanisms of research funding. The United Kingdom has supported one of the pioneering efforts to map the genome of a nonhuman organism and additional work to develop new mapping and sequencing technologies. Italy has the most specific commitment to the human genome: It funded several pilot projects (up to $1 million per year for 2 years) to map and perhaps sequence at least one small human chromosome, with the intent of increasing that budget five- to ten-fold if the projects are promising. France, the Federal Republic of Germany, and other Western European nations have substantial commitments to genetics research and are also discussing international cooperation. Canada's medical research planning board is considering special efforts for genome projects. The European Molecular Biology Laboratory and European Molecular Biology Organization have expressed interest in an international collaboration to map and sequence the genomes of human and nonhuman organisms.

Eastern European and Asian nations have expressed interest in using the resulting data, but they have relatively limited programs for genetics research. Australia is one possible exception; it has consistently increased its share of publications related to genetics over the last decade, and it would logically be included in any international planning. Japan is another exception. Its Science and Technology Agency has expended $3.8 million to support automation of DNA sequencing technologies, the Ministry of Education supports a grants program in genetics, and the Ministry of International Trade and Industry has devoted several million dollars to study the feasibility of an expanded international effort called the Human Frontiers Science Program, which could include genome research projects.

MISPLACED CONTROVERSY ABOUT "THE HUMAN GENOME PROJECT"

Over the past several years, the debate about genome projects has been vigorous—sometimes acrimonious. Many articles have appeared in the scientific press and the general press about "the genome controversy." The most conspicuous disagreements, however, have concentrated on issues that are not central to the conduct of genome projects. Disputes among the executive agencies have been played up, belying the generally close cooperation among DOE, NIH, HHMI, private firms, and other groups in conducting their respective projects. International cooperation among gene mappers and database managers has been successful but has attracted little attention. Private corporations are already involved in many of the projects that are furthest along. One firm has developed an extensive map of human genetic markers, and others have developed instrumentation useful in research relevant to the mapping and sequencing of DNA. These companies have offered few complaints about barriers to technology transfer. **Dissent has focused on the importance of and strategy for sequencing DNA of the entire genome, yet no agency has made a commitment to massive sequencing**. The current commitment is to develop *technologies* that would make it faster and less costly and to improve databases to collect and disseminate the resulting information. DOE has expressed interest in a concerted sequencing program, but only when technological development reduces its cost to tens of millions of dollars, in several years at the earliest.

Some of the debate can be attributed to the title that has often been applied to genome projects —the Human Genome Project. The term is a useful way to link research initiatives and to distinguish them from ongoing programs for budget planning. It highlights the *ultimate* objective—understanding human biology by developing a new set of research resources—and captures political support and broad public interest. It has had the effect, however, of generating rancorous debate which has inhibited the development of consensus on how to improve the research infrastructure. The importance of maps, databases, and repositories has been obscured by the controversy over massive DNA sequencing.

The title has had several other untoward effects. The Human Genome Project centers attention exclusively on human genetics, but **understanding human genes will necessarily involve the study of other organisms**. Many of the resources—particularly maps of human chromosomes—*will* be focused on human beings; but to interpret human genetic information, similar resources must be developed for other organisms. New instruments and methods will be applicable to *all* DNA.

The Human Genome Project invites confusion by implying that the human genome will be understood when the project is over. The immediate goal of genome projects is not complete understanding, but creating tools to bring about such understanding in the 21st century. Understanding encompasses all biomedical research; it does not distinguish genome projects from others. The most ambitious possible goal of genome projects would be to complete the most detailed map: a reference sequence of the entire human genome. Even if this were agreed to and developed, it would not yield immediate understanding of how that DNA sequence is translated to make a human being. It would not explain how nerve cells become connected in the immensely complex anatomy of the brain. It would not even provide complete answers to how individuals differ or how they have evolved. Sequence data, like other genetic information, is meaningful only when compared among individuals and correlated with biological function.

There is no single, monolithic Human Genome Project. In fact, there are several distinct components at various stages of development. Some instruments and many databases already exist; some genetic maps are more than half complete; repositories for DNA used in research have only been organized in the past few years; and other projects are planned but not yet begun. Whether there will ever be large and expensive research facilities for component specific genome projects is an open question that can be answered only as the technologies evolve.

The Human Genome Project conjures up images of large-scale projects such as the Manhattan Project to build the first atomic bomb, the Apollo Project for a manned Moon landing, the space station, or the superconducting supercollider. Genome projects do not belong in this category. Component genome projects will not require budgets as large as such megaprojects, nor are the technical ends as focused. Genome projects must be distinguished from the sequencing of the entire human genome, which is but a component still in the planning phase. There will be no single event such as the Moon landing or the space shuttle launching, nor is there likely to be construction of a new multi-billion-dollar facility such as the superconducting supercollider. Genome projects do not now require such facilities. Some projects may require facilities to perform services for mapping or sequencing in the future, yet such facilities would not be larger than the molecular biology centers already established at a few major research universities. Mapping or sequencing facilities would differ only by being devoted to production work rather than pure science. The results of genome projects are not contingent on completion of large capital-intensive dedicated units, and the data and instruments will be integrated into biology and medicine as the projects progress. Genome projects are, in this respect, analogous to navigational charts or road maps, which are useful even as they are being updated. Some persons believe a shortage of trained scientific and technical personnel in the United States could prove troublesome for molecular biology, but the genome projects proposed thus far are not so large in scale, even in comparison to other areas of biology, as to cause shortages in other areas. Genome projects are relatively modest compared to other large science projects now under consideration by the Federal Government.

THE CORE ISSUE: RESOURCE ALLOCATION FOR RESEARCH INFRASTRUCTURE

Most issues that need to be addressed regarding genome projects are variations on the problem of the commons: how to create and maintain resources of use to all. It can be difficult to develop goods useful to all if each individual has no direct incentive to pay for them and only a few are adversely affected.

The core issue concerning genome projects is resource allocation. What priority should be given to funding databases, materials repositories, genetic map projects, and development of new technologies? Should genome projects have precedence over other projects important to biological and biomedical research? These projects will benefit the entire biomedical research community, and ultimately the Nation and the world, but their funding must be drawn from the same agencies that support basic research. Funding for genome projects will thus be taken from agencies that support research on neuroscience, cancer, immunology, and many other promising and rapidly moving fields.

The flow of information from molecular biology is overwhelming the resources devoted to handling it. Federal agencies, HHMI, and other interested groups are acting to manage the deluge. **Research dedicated to improving databases, maps, repositories, and research methods promises to increase efficiency overall by doing once systematically what would otherwise be duplicated by many groups using more primitive technologies**. Whether massive, concerted DNA sequencing is similarly efficient can only be demonstrated by trying it on a smaller scale.

ORGANIZATION OF THIS REPORT

The following sections describe options for congressional action. Subsequent chapters address the issues raised here in greater detail. Chapter 2 provides technical background and explains how genome projects might be conducted. Chapter 3 reviews how results might be used in biology and medicine. Chapter 4 outlines some long-term social and ethical issues surrounding human genome projects. Chapter 5 surveys agencies and organizations in the United States actively supporting human genome projects. Chapter 6 discusses how genome projects might be organized among these agencies and organizations. Chapter 7 briefly surveys activities in foreign countries, and chapter 8 presents issues involved in technology transfer. Appendixes contain background on material used to produce this report, databases, costs of projects, and mapping and sequencing publications.

THE ROLE OF CONGRESS

Genome projects have come to the attention of Congress for three reasons. First, they have become highly visible because of the extensive debate surrounding them. Second, they involve agencies in different executive departments; therefore, mechanisms for coordinating them are less clear than if they were all in a single department. Third, results of genome projects will lead to new scientific and medical instruments for analysis of DNA, development of new genetic tests for use in clinical diagnosis, and other products and services. Techniques developed to analyze DNA will expedite biological research and will provide data and technologies crucial to the development of many new products. In this sense, genome projects promise economic returns, although the form and magnitude of them are not predictable. Genome projects have thus been linked to international competitiveness in biotechnology and its economic implications for American commerce in coming decades.

Congress has three roles regarding genome projects:

1. annual *appropriations* to Federal agencies funding the projects;
2. *authorization* of actions by executive agencies to set up formal coordinating structures or of specific mandates of agencies; and
3. *oversight* of agencies' conduct of their projects.

OPTIONS FOR ACTION BY CONGRESS

Options for congressional action discussed here build on the discussions above and those in chapters 4 through 6. Background material and details can be found in those chapters.

Appropriations to Federal Agencies

The pace of federally funded genome projects will be determined principally by the annual appropriations set by Congress and by the executive agencies' commitment to the projects. Although agencies retain some authority to "reprogram" funds for activities that fall within their mandates, large efforts cannot be sustained without specific appropriations. Appropriations will set an upper limit on the size and number of projects that are federally supported; commitment by executive agencies, and their grantees and contractors, will determine the speed and scope of projects within those limits.

The critical judgment in appropriations is the importance of the work to be supported relative to other research and activities supported by the Federal Government. The two national scientific groups that have written reports on genome projects, a DOE advisory subcommittee and an NRC committee, have both recommended substantial additional funding for genome projects, eventually equaling $200 million per year. OTA independently projected costs of genome projects at a workshop and through subsequent interviews and letters. Appendix B summarizes cost estimates, including the history of those made by other groups, and reviews the process OTA used to make its estimates. The cost of funding all component projects was estimated as increasing from $47 million the first year to $228 million the fifth year. This would permit strengthening of databases and repositories, construction of several varieties of chromosomal maps, development of many new technologies, and initiation of pilot projects for DNA sequencing.

Access to Information and Materials

The information produced by genetics research has swamped existing management systems. Materials to facilitate molecular genetic research have also proliferated, straining the resources devoted to making them widely available. These management problems will intensify as new technologies further accelerate research. Several of the genome projects are intended to systematically archive information, collect and store research materials, and make information and materials widely available to the research community. **Improving database and repository services is imperative whether or not other genome projects proceed**. If genetic mapping and sequencing initiatives are pursued, then databases and repositories will be needed even more. Bills have been introduced to improve coordination of and access to molecular biology databases through the National Library of Medicine. Each major repository and database has its own advisory panel of outside scientists. NIH has appointed an internal committee to report on NIH-supported repositories. Two international meetings were held in 1987 to discuss management of databases that contain DNA sequence data. NIH and DOE cosponsored a meeting on databases and repositories in August 1987, and appropriations to DOE and NIH have been increased to support databases and repositories. Congress has the options of maintaining current funding levels or increasing funds for database and repository services through the current system of agency planning and congressional oversight. Seeking recommendations from an advisory committee on how to integrate the development of databases and repositories with genome projects is an additional option.

Organization of Genome Projects

Congress could pass legislation to organize human genome projects—in fact, bills on organization have dominated discussion in Congress. There are four principal choices: 1) to designate a single agency to coordinate the projects, 2) to establish an interagency task force, 3) to establish a national consortium, or 4) to rely on congressional oversight of interagency agreement and consultation.

Establishing an interagency task force through legislation or encouraging agencies to do so by oversight are the least problematic choices. Designating a lead agency would be politically troublesome and would risk interruption of ongoing research programs at one or more agencies. Devising a single national consortium to manage the many diverse genome projects is likely to prove impractical. See chapter 6 for a more detailed discussion of these options.

Designate a Lead Agency

Congress could choose to designate a lead agency to coordinate and provide principal funding for genome projects. The chief advantage of a lead agency is accountability through clear authority. The purpose of focusing authority would be to reduce duplication of effort, to enhance coordination, and to give Congress a single agency on which to concentrate oversight. The chief disadvantage is that the difficult political process of selecting a lead agency would delay progress and diminish overall funding. If line item funding for genome projects at the nonlead agency—NIH or DOE—were eliminated, then agreement would have to be reached to add funds for the lead agency. This is a difficult process because it involves a completely different set of congressional committees and subcommittees for each agency. The choice of a lead agency would likely precipitate a protracted battle among agencies and congressional committees, which could only serve to delay projects. Furthermore, activities of NIH, DOE, NSF, HHMI, and other organizations are complementary rather than competitive and duplicative. Appointing a lead agency could complicate planning for the other agencies. As an alternative, each agency could take the lead in projects best suited to its mandate and expertise. This would result in a task force or consultative arrangement, discussed below, rather than a single lead agency. Designating a lead agency would attempt to centralize authority, but it is not clear that this would improve efficiency, communication, or coordination.

Designation of a lead agency for genome projects could, paradoxically, diminish rather than enhance accountability to Congress. This follows from the organizational structure of congressional committees. Genome projects supported by NIH, DOE, and NSF are authorized by several committees and subcommittees in both the House of Representatives and the Senate. Currently, each committee or subcommittee has independent authority to oversee programs in agencies under its jurisdiction, and interest in human genome projects has been high. Designating a lead agency would limit most oversight responsibility to a single committee. Further, a lead agency could not fully centralize authority, because HHMI is a nongovernment organization. Picking a lead agency would be politically difficult and is unlikely to occur unless there is strong evidence of the advantages of centralized authority for Federal efforts. The evidence to date is quite to the contrary: Agencies are communicating, sharing personnel, using compatible peer review procedures, and jointly funding projects in overlapping areas.

Designating a lead agency might eliminate pluralism in Federal funding of genome projects. An investigator wishing to pursue a genome project can now apply to NIH and DOE, or NIH and NSF for funding (depending on the nature of the project). If there were a single lead agency controlling genome projects, the choices would be limited, diminishing the pluralistic funding that has been a mainstay of American biology. If the lead agency had only an administrative role and did not provide the greatest amount of funds, then there would be little point in calling it a lead agency.

Congress sets independent budgets for NSF, NIH, and DOE through different subcommittees in the House and Senate appropriations committees. With several subcommittees involved, projects have alternative sources of support in Congress. Designating a lead agency would reduce this flexibility. The danger of pluralism is that different agencies will duplicate each other's work, will fail to cooperate, will fail to identify gaps in research, or will receive uncoordinated or inappropriate appropriations due to the absence of a clear authority structure. To date, such funding disarray has failed to materialize. There are checks and balances in the congressional budget process, through the Office of Management and Budget (OMB), and through the interagency consultation group in OSTP.

Arguments for a centralized and highly organized effort would be stronger if genome projects addressed a national health emergency, such as AIDS or polio, or if they were aimed at a single technical or scientific objective. But genome projects are many and diverse. Focused responsibility may nonetheless become necessary for some of them. Mapping, for example, might be more efficiently done at production centers as methods mature, and DNA sequencing might require dedicated facilities if the technology demands high capital investment or central management. If dedicated service centers are established, administration by a single agency or formal interagency agreement would be necessary to ensure standardization and efficiency. Such services would only be components of overall genome projects, however; integration of the various projects would still be needed.

If genome projects were neglected or inconspicuous elements in agencies' programs, then the advantages of central oversight through a single agency would carry more weight. This has not been the case. Genome projects have been given high priority—first by DOE and more recently by NIH—and there has been extensive media attention to agencies' management of them. There is thus little danger in the foreseeable future that genome projects will receive insufficient attention or that mismanagement will escape congressional scrutiny.

The agency most affected by genome projects will be the NIH. If Congress finds that the advantages of a lead agency outweigh the disadvantages, then NIH is the natural choice for lead agency. This is because biomedical research is NIH's central mandate, whereas NSF's and DOE's research programs include physical as well as life sciences. NIH funds over 10 times more genetics research than any other government or nongovernment organization, and researchers funded by NIH are the most numerous of the intended beneficiaries of genome projects. Researchers supported by DOE, NSF, and other organizations have important contributions to make, however, and some projects fall outside the mainstream of research supported by NIH. Genome projects that involve expertise in physical science, engineering, and other fields outside biomedical research would benefit from participation in or leadership by NSF or DOE. DOE in particular has vigorously promoted a Federal program to develop new technologies and to create sets of ordered DNA fragments. Some DOE-supported projects are logical extensions of work at the national laboratories, and DOE is the natural agency to conduct these. If NIH were designated the lead agency, it would be important to recognize and plan for the ongoing efforts of DOE.

Establish an Interagency Task Force

The chief advantage of an interagency task force is that it builds on existing research programs and planning efforts in different agencies and does not require a single lead agency. A task force could monitor all genome projects, government and nongovernment, obtain scientific advice, foster communication, and make recommendations to Congress and the appropriate agencies. Discussion at an OTA workshop in August 1987 stressed that agencies should have outside scientific advice and that advice given to one agency should take into account activities supported by other agencies. No advisory body exists to carry out this task. The chief disadvantage of a task force is that no one agency is accountable for the conduct of genome projects.

Creating a task force entails decisions about who should be represented, how appointments are to be made, and where the task force would be located administratively. Legislation could specify that it represent government, academic, industrial, and other relevant expertise and could stipulate the terms of membership and the appointment process. The task force could be made part of a government agency (making it in effect the lead agency), administratively autonomous, or attached to an existing quasi-governmental institution such as the National Academy of Sciences. Several bills to establish such coordination and advisory groups have been introduced in the 100th Congress and are likely to be acted upon in 1988.

Create a National Consortium

A consortium would involve one or more agencies in concert with private partners to support genome projects. The chief advantages of a consortium are administrative flexibility, possible funding by private firms to reduce government funding, and direct involvement of industrial partners—which would presumably hasten technology transfer. Some potential disadvantages are unclear lines of authority (caused by competing needs of government and private partners) and statements by the private sector that genome projects should be funded exclusively by the Federal Government (e.g., a poll taken by the Industrial Biotechnology Association). Accountability would be complicated in two respects. First, there are many genome projects, and it is difficult to imagine a single consortium that could oversee them all. Second, the possible commingling of government and nongovernment funds could prove troublesome. Consortia might nonetheless be formed for specific tasks. Some genome projects in technology development will undoubtedly be of great interest to industry and might attract private funding. Such projects (e.g., developing automated DNA mapping instruments or DNA detection methods) are likely to be highly focused, however, and organized at the local rather than the national level. Accountability would not be as diffuse for local consortia focused on specific technical objectives as for a single national consortium with multiple objectives and dozens of projects to manage.

The Technology Transfer Act of 1986 (Public Law 99-502) grants government agencies authority to form consortia with private corporations and provides guidelines for doing so. President Reagan's Executive Order 12591 (April 1987) further extends this authority and encourages federally owned laboratories to form consortia. Agencies thus have the requisite authority already. If Congress finds terms of the 1986 bill inappropriate in some details—for example, regarding patent policies or royalty arrangements—then the statute could be amended or special measures relating to genome projects could be added as amendments to other bills.

One bill introduced early in the 100th Congress would have established a national consortium specifically to manage genome projects, but the bill has since been replaced by one that establishes a new advisory body (covered above as a task force). A national consortium is not the only, and perhaps not the most effective, way to obtain industrial input for genome projects and to facilitate technology transfer. Alternatives are to encourage agencies to participate in the formation of local consortia; to facilitate exchange of industrial and academic expertise through training exchange programs, symposiums, and other mechanisms; and to include industrial representation on any national advisory groups.

Rely on Congressional Oversight

If Congress takes no explicit action, several outcomes are possible. Federal agencies could continue planning processes similar to those followed in 1986 and 1987, consisting of informal communication and coordination through an interagency group with members from NIH, DOE, NSF, OSTP, OMB, and other agencies. To date, NIH, DOE, and NSF have sought outside advice from various standing advisory committees, a practice that has resulted in conflicting recommendations. This problem could be remedied without legislation: The agencies could establish a single interagency advisory committee of outside experts appointed by the agencies or by a third party, such as the National Academy of Sciences or a private philanthropy. The advisory committee could report to the agencies directly.

A Committee on Life Sciences is forming in OSTP. The interagency nature and conspicuousness of genome projects make them a natural topic for this committee. OSTP is considering the creation of a special subcommittee on genome projects. Whether OSTP's efforts meet the objectives desired by Congress will depend on effective coordination and an appropriate balance among government, university, industrial, philanthropic, legal, bioethical, and other representatives on the subcommittee. If OSTP's subcommittee is composed exclusively of government representatives, then its primary function will be interagency communication. The main stumbling block to interagency planning to date has been conflicting advice from outside advisory bodies, not lack of interagency communication. Pluralism in funding is usually a virtue, but making conflicting recommendations to different agencies is not. Any national coordinating group should take a global view of activities in all agencies and harmonize the advice given them.

The chief advantage of relying solely on congressional oversight is that it requires no new legislation. One disadvantage is that interagency agreement on appointments and operating budgets for a coordinating body might prove difficult without a congressional mandate and might not initially include an appropriate range of nongovernment experts. Another potential disadvantage is that initiatives undertaken by an administration in the absence of legislation could crumble under the weight of later interagency disagreements or neglect by a subsequent administration. Flexibility is beneficial if projects are short-lived, but genome projects are not. Long-term stability is essential to the efficient conduct of genome projects because they will require sustained support over many years. Oversight of agency action could nonetheless be all that is required. Deficiencies of a task force set up by agencies could later be modified indirectly through congressional oversight or threat of legislation.

Technology Transfer

Congress appropriates funds to support scientific research for several reasons, the principal one for biomedical research being to improve health. Increasingly, however, biomedical research is being regarded as a national investment, and policies to facilitate economically fruitful applications of new knowledge are receiving attention in Congress. The process of exploiting new knowledge for practical purposes is called *technology transfer*. Some persons favor increased funding for genome projects because they believe the projects will lead to marketable products (instruments, research materials) or will accelerate research in areas that will later yield marketable products. Technology transfer can be improved through patent policies, exchange of industrial and academic personnel, symposiums for industrial and academic scientists, formation of consortia to develop specific technologies or services, and engaging industry in planning genome projects. Programs for exchanging personnel and sponsoring symposiums will fall to agencies through normal policy paths and can be monitored by Congress. Consortium formation and industry representation on planning bodies have been discussed above. The remaining policy area is patent and copyright law.

Patent policies of Federal agencies have changed dramatically during the past decade. The Patent and Trademark Amendments of 1980 (Public Law 96-517), as amended in 1984 (Public Law 98-620), were devised to facilitate commercialization of federally sponsored research. President Reagan issued directives to Federal agencies in February 1983 and April 1987 to this same end. And Congress passed the Technology Transfer Act of 1986 (Public Law 99-502), which contains patent licensing and joint venture provisions with authority to form consortia with private interests. These patent policies, following outlines of policies pioneered by NIH and NSF in the late 1970s, encourage institutions receiving Federal grants or contracts to patent products and processes resulting from federally funded work. A 1987 General Accounting Office report judged that the policies have increased patenting of research results.

Aside from a possible change regarding DOE policies (see ch. 8), genome projects raise no new questions of patent or copyright law. Genome projects would be subject to the same statutes and executive orders as other scientific efforts. There is a clear role for congressional oversight, however, in ensuring that data are shared promptly and fully.

In mid-1987, proposals to form private corporations to map and sequence the human genome stirred a controversy. Scientists expressed concern that scientific exchange would be impeded by such efforts and that information would be sequestered through copyrights and patents. If private corporations do form to develop map and sequence data and research materials, they will operate at private expense. If they are successful, scientists will have new information, services, and materials available for a price. If they fail, scientists should be no worse off, unless the government fails to support work it would otherwise have funded. To date, government agencies have not dropped plans for genome projects because of corporate efforts.

Corporate efforts need not entail restricted access to information. Corporations can provide services not appropriately performed by laboratories conducting basic scientific research (e.g., mapping, sequencing, or database management). Universities and large corporations can manage research facilities, such as the national laboratories, under contract. Corporations could also participate in consortia focused on specific technical objectives. Private firms could be given grants to develop new methods under the Small Business Innovation Research program; they would retain title to inventions, but they would have the same obligation to share data and materials as universities or other grantees. The essential point is not whether a grantee or a contractor is a university or corporation, but whether the research results will be widely shared.

It is essential to ensure timely exchange of data and materials from federally sponsored projects. Maps, databases, and repositories will be useful only if they are accurate and complete; they will be complete only if all participants make prompt contributions. In most cases, patent requirements should not substantially delay disclosure of data. Many data will not be relevant to a patentable invention. When research results do include a patentable invention, advance planning for filing patent applications should minimize delays. The main option for Congress in this area is to oversee the conduct of genome projects. Changes in agency policies for data exchange could be made if problems emerge.

Congress could also direct agencies to make it easier for persons receiving Federal grants or contracts to understand patent policies in the United States and abroad. At present, many of the published guidelines and regulations for NIH, DOE, and NSF are out of date. Investigators contemplating genome projects will probably contact more than one Federal agency for research support; it would be helpful to have a document summarizing the practices of the different agencies. Such a document could also explain the benefits of filing patents early and outline procedures for patenting abroad.

Questions for Congressional Oversight

Congressional oversight will most often involve an informal exchange among congressional staff, executive agency personnel, and other interested parties. Oversight can be a potent incentive for cooperation among agencies and for good conduct of executive actions. Congress may wish to hold hearings from time to time to address such questions as: Are genome projects being efficiently administered? Are agencies duplicating efforts on genome projects? Are agencies communicating effectively? Are agencies ensuring that access to shared data is relatively easy and fair? Are databases receiving the information they need to be most useful (e.g., map and sequence data)? Are commercial opportunities being exploited? Are shared research resources being neglected? Are issues of special interest to Congress, such as social and ethical implications of genome projects, being adequately addressed? Do genome projects supported by Federal agencies reflect national needs and social priorities? Are foreign governments funding a proportionate share of genetics research and the research infrastructure? Are foreign governments sharing data and materials to the same degree as U.S. agencies?

Chapter 2

Technologies for Mapping DNA

CONTENTS

Page
Organization and Function of Genetic Information 21
  What Is a Genome? 21
  How Are Genomes Organized? 21
  What Is the Genetic Code? 21
  How Big Is the Human Genome? 24
  How Does the Human Genome Compare to Other Genomes? 24
  Why Does Hereditary Information Change? 25
Genetic Linkage Maps 26
Linkage Maps of Restriction Fragment Length Polymorphisms 28
  RFLP Mapping 28
  When Is a Map of RFLP Markers Complete? 29
Low-Resolution Physical Mapping Technologies 30
  Somatic Cell Hybridization 30
  Chromosome Sorting 31
  Karyotyping 32
  Chromosome Banding 33
  In Situ Hybridization 33
  Other Methods for Mapping Genes 34
High-Resolution Physical Mapping Technologies 35
  Cloning Vectors as Mapping Tools 35
  Physical Mapping of Restriction Enzyme Sites 37
  Physical Mapping of Nonhuman Genomes 41
  Strategies for Physical Mapping of the Human Genome 43
  DNA Sequencing Technologies 44
Automation and Robotics in Mapping and Sequencing 46
Chapter 2 References 48

Figures

Figure *Page*
2-1. The Structure of DNA 22
2-2. Replication of DNA 22
2-3. Gene Expression 23
2-4. Number of Human Gene Loci Identified From 1958 to 1987 24
2-5. Separation of Linked Genes by Crossing Over of Chromosomes During Meiosis 26
2-6. Detection of Restriction Fragment Length Polymorphisms Using Radioactively Labeled DNA Probes . . . 29
2-7. Somatic Cell Hybridization 31
2-8. Chromosome Purification by the Flow Sorter 32
2-9. DNA Cloning in Plasmids 36
2-10. Separation of Intact Yeast Chromosomes by Pulsed Field Gel Electrophoresis 38
2-11. Constructing a Library of Clones Containing Overlapping Chromosomal Fragments 38
2-12. Making a Contig Map 39
2-13. DNA Sequencing by the Sanger Method 45
2-14. DNA Sequencing by the Maxam and Gilbert Method 46
2-15. Automated DNA Sequencing Using Fluorescently Labeled DNA 48

Tables

Table *Page*
2-1. The Genetic Code 23
2-2. Haploid Amounts of DNA in Various Organisms 25
2-3. Amount of Genome Sequenced in Several Well-Studied Organisms 47

Chapter 2

Technologies for Mapping DNA

ORGANIZATION AND FUNCTION OF GENETIC INFORMATION

What Is a Genome?

The fundamental physical and functional unit of heredity is the *gene*. *Genetics* is the study of the patterns of inheritance of specific traits. The chemical bearer of genetic information is *deoxyribonucleic acid (DNA)*. The DNA of multicellular organisms such as insects, animals, and human beings is associated with protein in highly condensed microscopic bodies called *chromosomes*. A single set, or *haploid* number, of chromosomes is present in the egg and sperm cells of animals and in the pollen cells of plants. All body cells, or *somatic* cells, carry a double set, or *diploid* number, of chromosomes, one originating from each parental set. **The entire complement of genetic material in the set of chromosomes of a particular organism is defined as its *genome*.**

How Are Genomes Organized?

Long before genetic material was identified as DNA, maps of genes on chromosomes were constructed, and many of the details of transmission of genes from generation to generation were elucidated [Judson, see app. A]. The gene for colorblindness, for example, was assigned to the human X chromosome in 1911 (80), about 40 years before the discovery of the structure of DNA. In fact, it has been known for nearly a century that the genetic material:

- has a structure that is maintained in stable form,
- is able to serve as a model for replicas of itself,
- has an information code that can be expressed, and
- is capable of change or variation.

Each of these features can be described in molecular terms based on the structure and function of DNA.

To know how DNA controls cell function, and ultimately the structure and function of an entire organism, it is necessary to understand its structure. In multicellular organisms, DNA is generally found as two linear strands wrapped around each other in the form of a double helix. A DNA strand is a polymeric chain made of *nucleotides*, each consisting of a nitrogenous base, a deoxyribose sugar, and a phosphate molecule (figure 2-1). The arrangement of nucleotides along the DNA backbone is called the *DNA sequence*. There are four nucleotides used in DNA sequences: adenosine (A), guanosine (G), cytidine (C), and thymidine (T). The two strands of DNA in the helix are held together by weak bonds between *base pairs* of nucleotides. In nature, base pairs form only between A's and T's and between G's and C's. **The size of a genome is generally given as its total number of base pairs.**

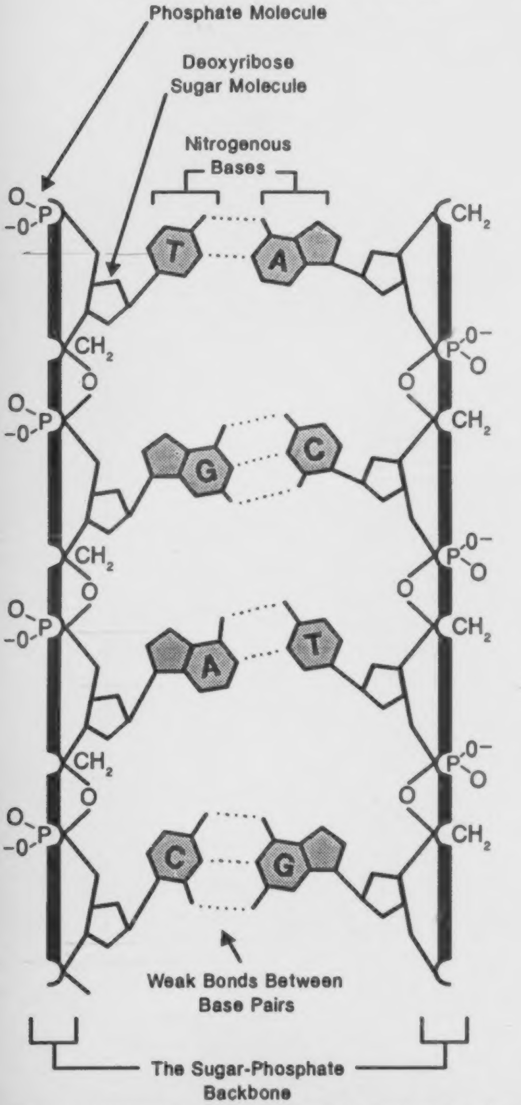
A full genome of DNA is regenerated each time a cell undergoes division to yield two daughter cells. During cell division, the DNA double helix unwinds, the weak bonds between base pairs break, and the DNA strands separate. Free nucleotides are then matched up with their complementary bases on each of the separated chains, and two new complementary chains are made (figure 2-2). In human and other higher organisms, DNA *replication* occurs in the nucleus of the cell. This DNA replication process was first proposed in 1953 by Francis H.C. Crick and James D. Watson (19,73,74).

What Is the Genetic Code?

Most genes carry an information code that specifies how to build *proteins*. Proteins are an essential class of large molecules that function in the formation and repair of an organism's cells and tissues. Proteins can be components of essential structures within cells, or they can carry out more active roles in the overall function of a particular cell type. Included in this important class of molecules are hormones such as insulin, antibodies to fight cellular infections, and receptors on the cell's surface for modulating interactions between a particular tissue and its surroundings (68). *Enzymes* are a specialized group of proteins that increase the rate of the biochemical processes that take place in metabolism.

Proteins are long chains of smaller molecules, called *amino acids,* that fold into the unique structures necessary for protein function. The information for generating proteins of specific amino

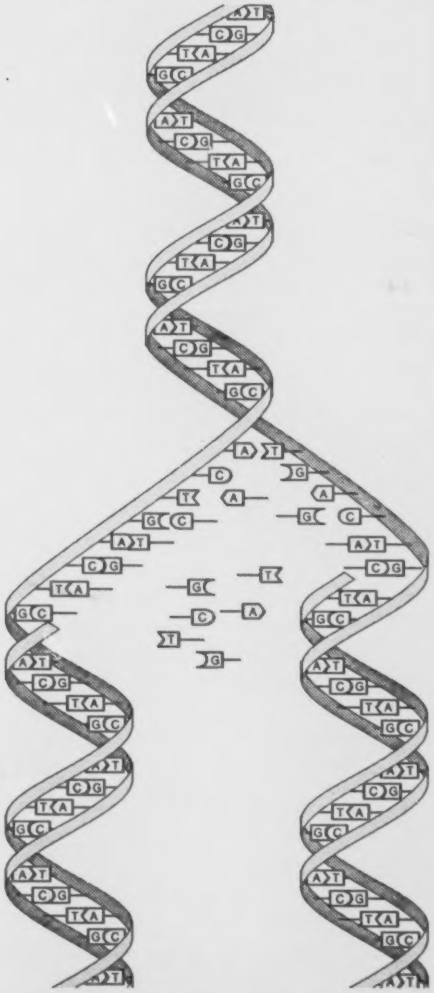
Figure 2-1.—The Structure of DNA


Phosphate Molecule
Deoxyribose Sugar Molecule
Nitrogenous Bases
T
A
G
C
A
T
C
G
CH2
O
P
-O
O-
Weak Bonds Between Base Pairs
The Sugar-Phosphate Backbone

The four nitrogenous bases, adenine (A), guanine (G), cytosine (C), and thymine (T), form the four letters in the alphabet of the genetic code. The pairing of the four bases is A with T and G with C. The sequence of the bases along the sugar-phosphate backbone encodes the genetic information.

SOURCE: Office of Technology Assessment, 1988.

Figure 2-2.—Replication of DNA



When DNA replicates, the original strands unwind and serve as templates for the building of new, complementary strands. The daughter molecules are exact copies of the parent, each daughter having one of the parent strands.

SOURCE: Office of Technology Assessment, 1988.

acid sequences is found in the *genetic code*—a code based on sequences of nucleotides that are "read" in groups of three (table 2-1). Genetic information is transmitted from DNA sequences to protein via another large molecule called *messenger ribonucleic acid (mRNA)*. The structure of *ribonucleic acid (RNA)* is very similar to that of DNA. Figure 2-3 illustrates the major steps in *gene expression*, namely:

Table 2-1.—The Genetic Code

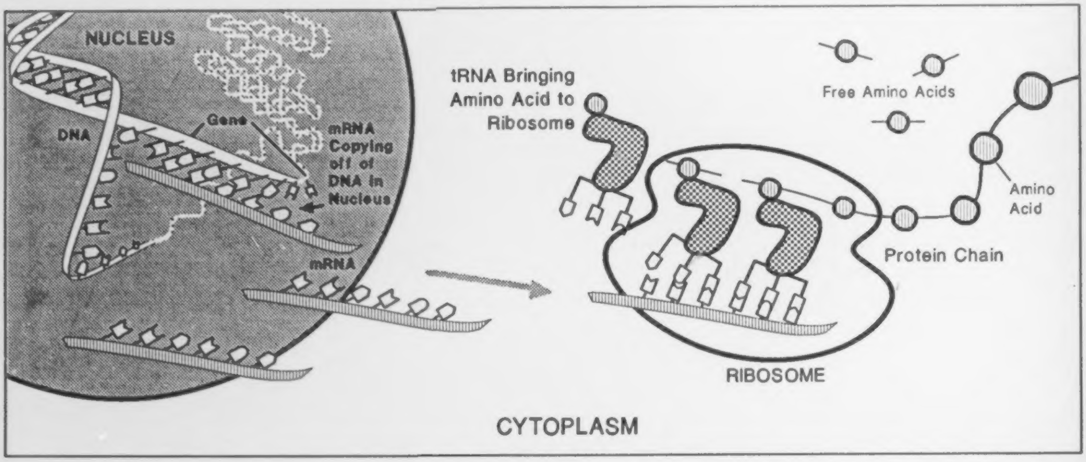
Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid	Codon	Amino Acid
UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
UUC	Phenylalanine	UCC	Serine	UAC	Tyrosine	UGC	Cysteine
UUA	Leucine	UCA	Serine	UAA	stop	UGA	stop
UUG	Leucine	UCG	Serine	UAG	stop	UGG	Tryptophan
CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine
CUC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine
CUA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine
CUG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine
AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
AUC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine
AUA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine
AUG	Methionine (start)	ACG	Threonine	AAG	Lysine	AGG	Arginine
GUU	Valine	GCU	Valine	GAU	Aspartic acid	GGU	Glycine
GUC	Valine	GCC	Alanine	GAC	Aspartic acid	GGC	Glycine
GUA	Valine	GCA	Alanine	GAA	Glutamic acid	GGA	Glycine
GUG	Valine	GCG	Alanine	GAG	Glutamic acid	GGG	Glycine

Each codon, or triplet of nucleotides in RNA, codes for an amino acid. Twenty different amino acids are produced from a total of 64 different RNA codons, but some amino acids are specified by more than one codon (e.g., phenylalanine is specified by UUU and by UUC). In addition, one codon (AUG) specifies the start of a protein, and three codons (UAA, UAG, and UGA) specify the termination of a protein. Mutations in the nucleotide sequence can change the resulting protein structure if the mutation alters the amino acid specified by a codon or if it alters the reading frame by deleting or adding a nucleotide.

U=uracil (thymine) A=adenine
C=cytosine G=guanine

SOURCES: Office of Technology Assessment and National Institute of General Medical Sciences, 1988.

Figure 2-3.—Gene Expression



In the first step of gene expression, messenger RNA (mRNA) is synthesized, or transcribed, from genes by a process somewhat similar to DNA replication. In higher organisms, this process takes place in the nucleus of a cell. In response to certain signals (e.g., association with a particular protein), sequences of DNA adjacent to, or sometimes within, genes control the synthesis of mRNA. Protein synthesis, or translation, is the second major step in gene expression. Messenger RNA molecules are known as such because they carry messages specific to each of the 20 different amino acids that make up proteins. Once synthesized, mRNAs leave the nucleus of the cell and go to another cellular compartment, the cytoplasm, where their messages are translated into the chains of amino acids that make up proteins. A single amino acid is coded by a sequence of three nucleotides in the mRNA, called a codon. The main component of the translation machinery is the ribosome—a structure composed of proteins and another class of RNAs, ribosomal RNAs. The ribosome reads the genetic code of the mRNA, while a third kind of RNA molecule, transfer RNA (tRNA), mediates protein synthesis by bringing amino acids to the ribosome for attachment to the growing amino acid chain. Transfer RNAs have three nucleotide bases that are complementary to the codons in the mRNA (see table 2-1).

SOURCE: Office of Technology Assessment, 1988.

- *transcription* of DNA into mRNA, and
- *translation* of mRNA into protein.

By these processes, the genetic code directs amino acids to be joined together in the order specified by the sequence of nucleotides in the messenger RNA, which was in turn determined by the sequence of nucleotides in the DNA.

In molecular terms, a gene is a region of a chromosome whose DNA sequence can be transcribed to produce a biologically active RNA molecule. Messenger RNAs constitute the major class of biologically active RNAs. Other RNAs may act as lattices to stabilize certain cell structures or may participate directly in important cellular processes such as protein synthesis.

How Big Is the Human Genome?

The diploid human genome consists of 46 chromosomes—22 pairs of *autosomes* and 1 pair of *sex chromosomes* (two X chromosomes for females and an X and a Y chromosome for males). A single egg cell has 22 different autosomes and a single X chromosome, whereas sperm cells carry 22 different autosomes and either an X or a Y chromosome.

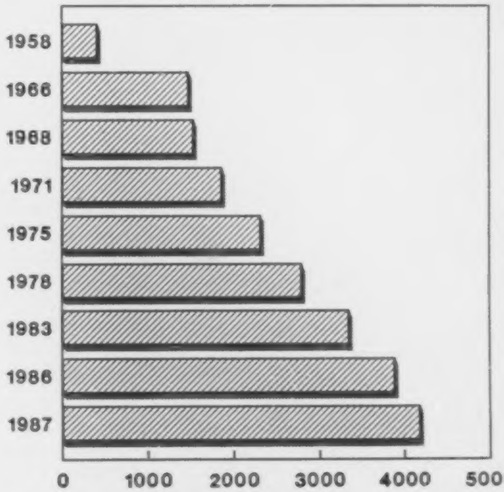
Scientists estimate the total number of human genes per haploid genome at 50,000 to 100,000. The characterization of the structure of human genes on chromosomes was made possible recently through *recombinant DNA technology* (the use of molecular biology tools to combine DNA from one organism with that of another). It is now known that human genes can vary in size from fewer than 10,000 base pairs to more than 2 million. The entire haploid genome is approximately 3 billion base pairs. So far, researchers are far from having determined where each human gene is located on the 24 chromosomes. Victor McKusick of The Johns Hopkins University maintains *Mendelian Inheritance in Man*, an encyclopedia of expressed genes [see app. D]. According to the October 1, 1987, count, 4,257 genes were represented in the encyclopedia; of those, at least 1,200 had been mapped to specific chromosomes or regions of chromosomes (51). Figure 2-4 illustrates the years of effort invested thus far in identifying even this small fraction of the total number of human genes.

How Does the Human Genome Compare to Other Genomes?

Before much was known about the DNA sequences that make up genomes, it was thought that the amount of DNA per haploid genome would increase in proportion to the biological complexity of the organism. Since chromosomes can vary in size, the total amount of DNA in a haploid cell is a better indicator of actual genome size than the number of chromosomes. Table 2-2 shows that higher plants and animals do have much more DNA than lower organisms. There are some notable exceptions, however, to the correlation between overall genome size and complexity of the organism. A good example is the salamander, which has a haploid DNA content more than 30 times greater than that of humans, even though it is obviously a smaller, less complex organism. Similarly, the cells of some species of plants have a greater DNA content than human cells (72).

This inconsistency between DNA content and the apparent complexity of an organism is known to geneticists as the *C-value paradox* (C-value refers to the haploid genome size). A great deal of research has been devoted to determining the scientific basis for the C-value paradox. Variations

Figure 2-4.—Number of Human Gene Loci Identified From 1958 to 1987



SOURCE: Victor McKusick, The Johns Hopkins University Medical School, Baltimore, MD.

Table 2-2.—Haploid Amounts of DNA in Various Organisms

Organism	Number of base pairs (millions)
Bacterium	4.7
Yeast	15
Nematode	80
Fruit fly	155
Chicken	1,000
Human	2,800
Mouse	3,000
Corn	15,000
Salamander	90,000
Lily	90,000

SOURCES:

B. Alberts, D. Bray, J. Lewis, et al., *Molecular Biology of the Cell* (New York, NY: Garland Publishing, 1983).

C. Burks, GenBank®, Los Alamos National Laboratory, Los Alamos, NM, personal communication, March 1988

T. Cavalier-Smith (ed.), *The Evolution of Genome Size* (New York, NY: Wiley & Sons, 1985).

J. Darnell, H. Lodish, and D. Baltimore, *Molecular Cell Biology* (New York, NY: Scientific American, 1986).

in genome size usually arise from increases in the amount of DNA per chromosome, not from increases in numbers of chromosomes. The genomes of all higher organisms contain sequences of DNA that occur as large numbers of repeated units, either clustered in one chromosomal region or in regions dispersed throughout the entire genome. These repeated sequences contribute to wide variations in total DNA content among what are often closely related species.

In large genomes such as the human genome, *intron* sequences also contribute to size. Introns are DNA sequences occurring within the coding region of a gene. They are transcribed into mRNA, but are cut (spliced) out of the message before it is translated into protein. Introns can increase the number of base pairs in a gene by more than tenfold. Many genes also have long regions at their ends that are transcribed into mRNA but are not translated into protein. In addition, some protein-coding genes have given rise to *gene families* that make several closely related protein products. Other gene families consist of hundreds or thousands of closely related genes (72).

The untranslated sequences within or at the ends of genes, gene families, and moderately or frequently repeated DNA sequences between genes still do not account for all of the DNA in the genomes of higher organisms, nor for the variations in genome size among these organisms. Many scientists interpret these facts to mean that some fraction of DNA in the human genome is expendable; although there is little agreement on the size of this fraction, some believe it to be more than 90 percent of the genome (27,54). The implication of the C-value paradox, that much of human DNA is expendable, is one reason that some esteemed scientists do not favor a major effort to obtain a complete nucleotide sequence of the human genome. They believe time would be better spent identifying and understanding the function of gene products that contribute to the cellular processes leading to the development of an organism as complex as man (1). On the other hand, some scientists consider the C-value paradox to be one of the many mysteries that might be unraveled once entire genomes have been analyzed in greater detail.

Why Does Hereditary Information Change?

Hereditary variation is the result of changes occurring by *mutation*—a change in the sequence or number of nucleotides—which occurs during DNA replication. Mutations formed in sex cells are inherited by offspring, whereas those that occur in somatic cells remain only in the affected organism. Some diseases, such as certain human cancers, arise from factors in both of these categories. Mutations are also acquired by artificial means, such as exposure to chemicals or certain forms of radiation.[1] Such factors can cause a change in a single DNA base pair that may modify or inactivate a protein, if one is encoded in that region of the chromosome.

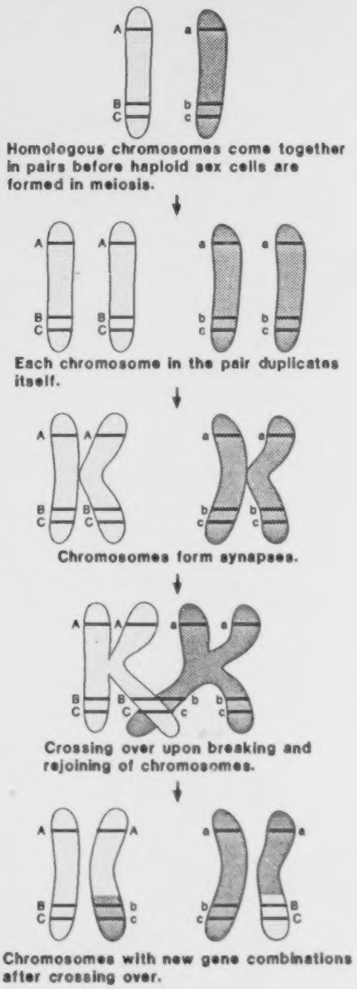
More extreme mutations, involving changes in the structure of a single chromosome or changes in chromosome number, can occur; for example:

- *deletion* of a chromosome,
- *duplication* of a chromosome or a piece of a chromosome,

---

[1] A 1986 OTA report, *Technologies for Detecting Heritable Mutations in Human Beings*, addresses the kinds and effects of mutations in human beings and new technologies for detecting mutations and measuring mutation rates.

Figure 2-5.—Separation of Linked Genes by Crossing Over of Chromosomes During Meiosis


A
B
C
a
b
c
Homologous chromosomes come together in pairs before haploid sex cells are formed in meiosis.
Each chromosome in the pair duplicates itself.
Chromosomes form synapses.
Crossing over upon breaking and rejoining of chromosomes.
Chromosomes with new gene combinations after crossing over.

SOURCE: Office of Technology Assessment, 1988.

- *translocation*, or insertion of a chromosome fragment from one chromosome pair into an unmatched member of a different pair, and
- *inversion*, or the breakage of a chromosome fragment followed by its rejoining in the opposite orientation.

In diploid cells, there is a tendency for each DNA molecule to undergo some form of modification or rearrangement with each cell division. The progenitors of sex cells are a special class of diploid cells that undergo two rounds of cell duplication in a process called *meiosis*. Meiosis results in four instead of two daughter cells, each with a haploid set of chromosomes. Before the first meiotic cell division, each member of a chromosome pair is replicated, forming two sets of chromosome pairs. At this stage, the cell has two identical copies of chromosomes of maternal origin and two identical copies of chromosomes of paternal origin. Also at this time, the chromosome pair of maternal origin is in close association with that of paternal origin, and an event called *crossing over* can occur; that is, one maternal and one paternal chromosome can break, exchange corresponding sections of DNA, and then rejoin (figure 2-5). (This process is also referred to as *recombination*.) In this way, two of the four resulting sex cells have chromosomes with new combinations of genes, while the other two cells carry the parental (original) combinations of genes. Since chromosomes originating maternally or paternally can carry different forms of any given gene, new combinations of traits are created by such crossovers.

GENETIC LINKAGE MAPS

Because of recombination during meiosis, certain groups of traits originating on one chromosome are not always inherited together (figure 2-5). The closer, or more *linked*, genes are on a particular chromosome, the smaller the probability that they will be separated during meiosis. Each chromosome is inherited independently of all others, so only genes on the same chromosome can be linked.

Gene mapping, broadly defined, is the assignment of genes to chromosomes. A genetic linkage map permits investigators to ascertain one genetic locus relative to another on the ba-

sis of how often they are inherited together. Strictly speaking, a genetic locus is an identifiable region, or *marker*, on a chromosome. The marker can be an expressed region of DNA (a gene) or some segment of DNA that has no known coding function but whose pattern of inheritance can be determined. Variation at genetic loci is essential to genetic linkage mapping. **The markers that serve to identify chromosome locations must vary in order to be useful for linkage studies in families, because only when the parents have different forms at the marker locus can linkage to a gene be followed in their children.** *Alleles* are the alternative forms of a particular genetic locus. For example, at the locus for eye color, there are blue and brown alleles. During meiosis, all of the genetic loci on a chromosome remain together unless they are separated by crossing over between chromosome pairs.

Distance on genetic maps is measured by how often a particular genetic locus is inherited separately from some marker. This measure of genetic distance is called *recombination frequency*. The amount of recombination is expressed in units called *centimorgans*. One centimorgan is equal to a 1 percent chance of a genetic locus being separated from a marker due to recombination in a single generation.

During the generation of sex cells in human beings, if a gene and a DNA marker are separated by recombination in 1 percent of the cases studied, then they are, on average, separated by 1 million base pairs. The relationship between genetic map distance (recombination frequencies) and physical map distance (measured in DNA base pairs) can vary, however, by five- or even tenfold. Recombination can vary from near zero, if genetic loci are very close, to 50 percent, between genetic loci that are far apart on the chromosome or on different chromosomes. Some chromosome regions are highly prone to recombination and exhibit high recombination frequencies, while other chromosome regions appear to be resistant to recombination. Interestingly, the rate of recombination in the same region of a particular chromosome typically varies among males and females, and it is often greater in females. The reasons for this have not been established. Double or multiple crossover events can also occur between two loci that are widely separated. Each of these variations in recombination frequencies complicates the relationship between genetic and physical maps. Nevertheless, if a genetic linkage map were constructed with a set of markers separated by an average of 1 centimorgan, then most genes could be located within a range of 100,000 to 10 million base pairs.

Genetic linkage between two or more observable traits can be established with greater certainty in large populations. For this reason, large families are preferred for mapping studies. If two genetic loci are closely linked, then their separation by recombination during meiosis is unlikely and a large family must be studied to determine how close they are on the genetic map. As more loci are placed on the genetic map, it becomes possible to determine the location of a new trait on the basis of its inheritance pattern compared to two or three others already on the map. The frequency with which multiple traits are inherited together generally must be calculated for many individuals over many generations before genetic mapping results are statistically significant.

The X chromosome is particularly amenable to linkage analysis because male traits directly reflect the genes on the single X chromosome present. For this reason, the genetic linkage map of the X chromosome is the most nearly complete of all chromosome maps.

Mapping of genetic loci on autosomes, on the other hand, is not as easy, unless the gene is found to be linked to a marker that has already been mapped through the study of family inheritance patterns. The first assignment of a gene to a specific autosomal chromosome came in 1968, when researchers showed that the Duffy blood group, which can be identified in families by biochemical methods, is linked to a variation in chromosome 1 (23). About the same time, the feasibility of correlating specific genes with particular chromosomes or chromosome regions by a technology called somatic cell hybridization was demonstrated (75). This and other experimental methods developed in the 1970s radically advanced the study of human genetics, allowing investigators to locate autosomal genes on human genetic and physical maps [Judson, see app. A] (50,58).

LINKAGE MAPS OF RESTRICTION FRAGMENT LENGTH POLYMORPHISMS

The advent of recombinant DNA technology in the 1970s brought about a tremendously useful new way to create genetic linkage maps. Examination of DNA from any two individuals reveals that variations in DNA sequence occur at random about once in every 300 to 500 base pairs (37). These variations occur both within and outside of genes, and most do not lead to functional changes in the protein products of genes. Kan and Dozy (40) were the first to demonstrate this phenomenon experimentally, by showing that one particular DNA sequence, recognized by the restriction enzyme HpaI, was lost in certain individuals. (*Restriction enzymes* are proteins that recognize specific, short nucleotide sequences and cut the DNA at those sites.) This alteration in the DNA correlated with the inheritance of sickle cell disease.

This important discovery led researchers to propose that natural differences in DNA sequence (*polymorphisms*) might replace other chemical and morphological markers as a way to track chromosomes through a family (5,64). In addition to polymorphisms in *restriction enzyme cutting sites,* it is possible to detect differences among individuals in the number of copies of short DNA sequences repeated in tandem. Polymorphic sequences can occur within a restriction enzyme cutting site or between sites. In either case, the lengths of DNA fragments generated upon cutting the DNA with restriction enzymes will vary among individuals having different alleles at such locations. These polymorphic sequences are thus commonly referred to as *restriction fragment length polymorphism (RFLP)* markers.

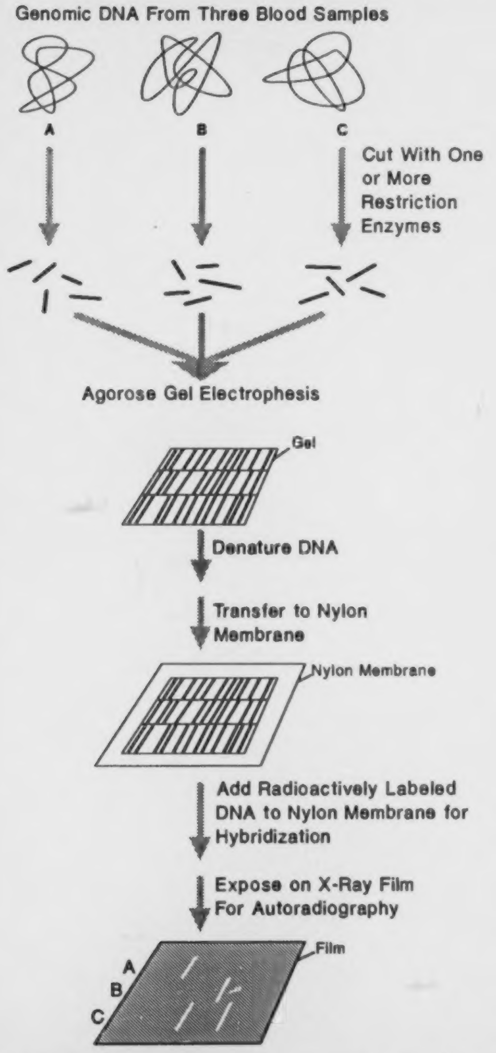
In 1983, genetic linkage between a RFLP marker on chromosome 4 and Huntington's disease (a neurological disease that usually strikes its victims by the age of 35) was discovered (31), paving the way for the general use of RFLPs as markers for genes responsible for inherited disorders. The more frequently a RFLP marker is inherited with the gene, the more likely it is to be physically close to the gene, and hence the more useful it is as a gene marker. The major limitations to the usefulness of RFLP markers are how polymorphic they are (how much they vary among individuals), how many other markers exist in the same region, and the extent to which DNA samples of large families are available for analysis (43,44).

RFLP Mapping

A RFLP map is a type of genetic linkage map, consisting of markers distributed throughout the genome. Construction of the map involves determining the linkages between RFLP markers, their arrangement along the chromosomes, and the genetic distances between them. RFLP markers are identified and mapped by comparing the sizes and numbers of restriction enzyme fragments generated from different individuals. Just as genetic loci representing expressed DNA segments have alternate, or allelic, forms, so may RFLPs. The value of any marker depends mostly on how many variants it displays. The more often the marker varies in a population, the more likely it is that an individual will inherit two different alleles at the marker location (one on each member of a matched pair of chromosomes), making it possible to detect recombination between markers in that individual's offspring (76).

In RFLP mapping, DNA obtained from white blood cells (lymphocytes) or other tissues of several different individuals are first cut into fragments using restriction enzymes (figure 2-6). The fragments are then separated by size. This is accomplished by a procedure called *electrophoresis*, in which a mixture of DNA fragments of various sizes is placed in a polymeric gel (e.g., agarose) and then exposed to an electric field. Because the chemical makeup of DNA gives it a net negative charge, the DNA fragments will travel in an electric field toward a positive electrode. Large DNA fragments will move more slowly than small ones, thus the mixture is separated, or resolved, according to size. With very large pieces of DNA, the use of restriction enzymes yields numerous fragments along the entire length of the gel, making it necessary to identify RFLPs using radioactively labeled, single-strand segments of DNA called *DNA probes* (65). RFLP markers are identified by virtue of their ability to form base pairs (hybridize) with DNA probes that have complementary sequences of nucleotides. Some useful probes for RFLP mapping are fragments of genes; others are randomly isolated DNA segments that identify polymorphisms; still others are complementary to sequences with variable numbers of tandemly repeated, shorter sequences that occur frequently within the genome. A technique called *autoradiography* is used to show the image of a band on an X-ray film wherever the agarose gel held a restriction enzyme fragment that hybridized to the DNA probe. Where polymorphisms occur, different patterns will be observed among samples taken from different individuals [Myers, see app. A] (figure 2-6).

Figure 2-6.—Detection of Restriction Fragment Length Polymorphisms Using Radioactively Labeled DNA Probes


Genomic DNA From Three Blood Samples
A
B
C
Cut With One or More Restriction Enzymes
Agorose Gel Electrophesis
Gel
Denature DNA
Transfer to Nylon Membrane
Nylon Membrane
Add Radioactively Labeled DNA to Nylon Membrane for Hybridization
Expose on X-Ray Film For Autoradiography
Film
A
B
C

Variations in DNA sequences at particular marker sites are observed as differences in numbers and sizes of DNA fragments among samples taken from different individuals (shown here as samples A, B, and C).

SOURCE: Office of Technology Assessment, 1988.

When Is a Map of RFLP Markers Complete?

Botstein and co-workers (5) predicted in 1980 that only 150 different markers would be needed to link all human genes to chromosomal regions containing RFLPs. In practice, however, it has been estimated that many more markers may have to be studied and evaluated in order to find the minimum number which would be randomly distributed over the genome (45). It now appears that hundreds of DNA probes for highly polymorphic sequences, scattered widely over the genome, will be required for a complete human linkage map (77).

With a 10-centimorgan map, for example, there is a greater than 90 percent chance of being able to determine the rough chromosomal location of any gene associated with an inherited disease. Raymond White and colleagues at the Howard Hughes Medical Institute at the University of Utah have taken advantage of one such tandemly repeated sequence, known as VNTR, to create a set of probes useful for making a complete RFLP map of the human genome (76). White's RFLP map, with continuously linked landmarks separated on average by 10 centimorgans (about 10 to 20 million base pairs), is nearly complete. At the ninth international Human Gene Mapping Workshop, held in September 1987, he reported 475 markers covering 17 human chromosomes, based on the DNA from 59 different three-generation families. White's group and other geneticists believe that a 1-centimorgan RFLP marker map, determined from normal families and consisting of thousands of markers spaced an average of 1 million base pairs apart, would be the ideal research tool (see ch. 3 for further discussion) (17,52).

Another group, led by Helen Donis-Keller at Collaborative Research, Inc. (Waltham, MA), reported its own RFLP linkage map, consisting of 403 markers an average of 9 centimorgans apart. A new gene or marker on their map can be located relative to the existing markers 95 percent of the time (24).

As physical markers that can be followed genetically, RFLPs are the key to linking the genetic and physical maps of the human genome. RFLP linkage maps, as well as linkage maps of expressed genes, can be correlated with banding patterns and other identifiable regions of chromosomes by somatic cell hybridization and *in situ* hybridization. These and other relatively low resolution physical mapping technologies are described in the following sections.

LOW-RESOLUTION PHYSICAL MAPPING TECHNOLOGIES

A physical map is a representation of the locations of identifiable landmarks on DNA. For the human genome, the physical map of lowest resolution is found in the banding patterns on the 22 autosomes and the X and Y chromosomes observable under the light microscope. This map has at most 1,000 landmarks (i.e., visible bands) (57).

Another type of relatively low resolution physical map illustrates the positions of expressed segments of DNA relative to certain regions of the chromosome or to specific chromosome bands. Expressed genes include those that are transcribed into mRNA and then translated into protein, and another class of essential genes that are transcribed into RNA but not translated into proteins. Included in the latter class are transfer and ribosomal RNAs involved in protein synthesis, RNAs involved in the removal of intron sequences from mRNAs, and an RNA associated with the cellular protein secretion machinery. Procedures are available to make DNA copies, or *complementary DNAs (cDNAs)*, of RNA transcripts. These cDNAs can in turn be mapped to genomic DNA sequences by somatic cell hybridization, *in situ* hybridization, and other low-resolution physical mapping methods. A physical map illustrating the location of expressed genes is often referred to as a cDNA map. As noted earlier, only 1,200 of the 50,000 to 100,000 human genes have been physically mapped to chromosomes.

A high-resolution physical map can be made by cutting up the entire human genome with restriction enzymes and ordering the resultant DNA segments as they were originally oriented on the chromosomes. This third type of physical map, a *contig* map, can be related to the maps of chromosome bands and expressed genes. **The physical map of highest possible resolution, or greatest molecular detail, is the complete nucleotide sequence of the human genome.** Thus there is a continuum of mapping techniques that ranges from low to high resolution (see table 2-2). These techniques are discussed in this section, on low-resolution physical mapping, and in the following one, on high-resolution physical mapping methods.

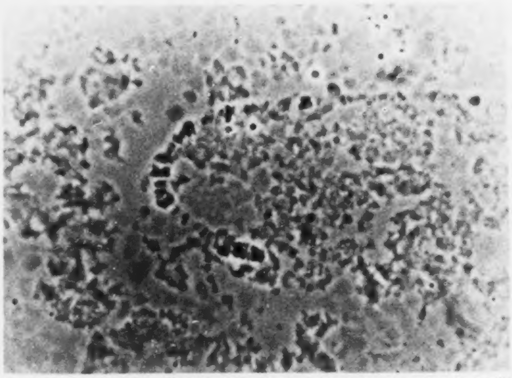


Photo credit: Stephen Mount, Columbia University, New York, NY

Banded pattern of *Drosophila melanogaster* salivary gland chromosomes as seen under phase contrast light microscopy.

Somatic Cell Hybridization

The somatic cell hybridization technique for gene mapping typically employs human fibroblast and rodent tumor cells grown in culture. The human and mouse cells are fused (hybridized) together using certain chemicals, Sendai virus, or an electric field, as illustrated in figure 2-7 (58). The chromosomes of each of the fused cells become mixed, and many of the chromosomes are lost from the hybrid cell. Human chromosomes are preferentially lost over rodent chromosomes, but there is generally no preference for which human chromosomes are lost. The individual hybrid cells are then propagated in culture and maintained as cell lines. In practice, the hybrid cell lines resulting from cell fusions contain different subsets of between 8 to 12 human chromosomes in addition to rodent chromosomes (58).

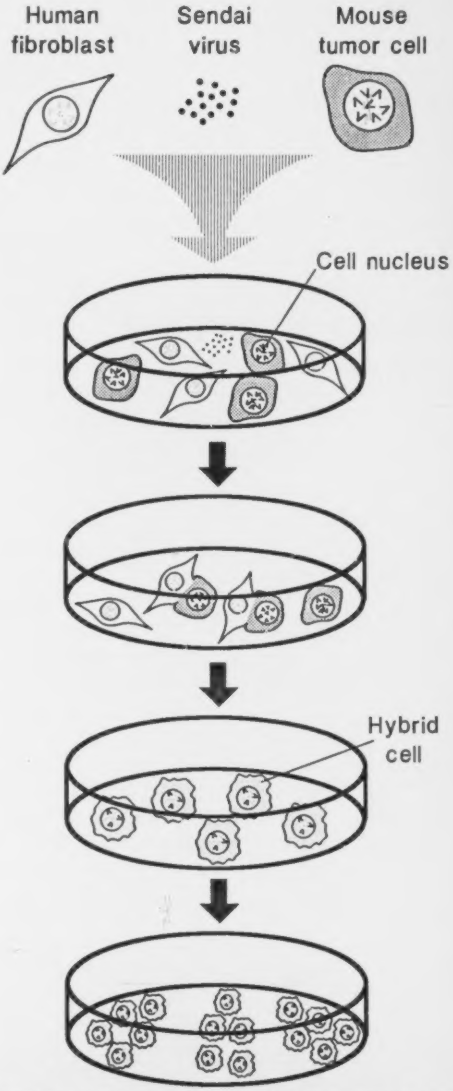
Using a large set (panel) of somatic cell hybrids containing different chromosome combinations, it is possible to correlate the presence or absence of a particular chromosome with a particular gene. Assignment of a gene to a chromosome is made by detecting a protein produced by a hybrid cell line and associating it with the chromosome unique to that cell line. Alternatively, if the gene to be mapped has already been isolated by DNA cloning procedures, then the gene can be used directly to identify complementary nucleotide sequences in the DNA extracted from somatic cell hybrids.

Modifications of the somatic cell hybridization method have been devised to generate single chromosome hybrids; to date, hybrid cells containing single copies of human chromosomes 7, 16, 17, 19, X, and Y are available (58). Somatic cell hybrid lines carrying chromosomes with deletion or translocation mutations are also useful low-resolution mapping tools because they make it possible to infer the location of a particular gene.[2]

Chromosome Sorting

Chromosome sorting offers an alternative to the screening of somatic cell hybrid panels for low-resolution gene mapping. In this approach, DNA hybridization is used to map genes to chromosomes that have been differentiated by flow

Figure 2-7.—Somatic Cell Hybridization



Somatic cell hybrids are generated by the process of cell fusion, an event that can be enhanced by adding Sendai virus. Initially, the hybrid cell contains complete sets of chromosomes from both parental cells, but hybrids of human and mouse cells are unstable and chromosomes from the human cells are preferentially lost. After a few generations in culture, a line of hybrid cells is established that contains both mouse and human chromosomes.

SOURCE: Office of Technology Assessment, 1988.

---

[2] The Institute for Medical Research, in Camden, New Jersey, established a repository for SCHs with chromosome rearrangements called the Human Mutant Cell Library. The availability of this centralized storage facility has accelerated the rate of mapping human genes to specific chromosomal locations.

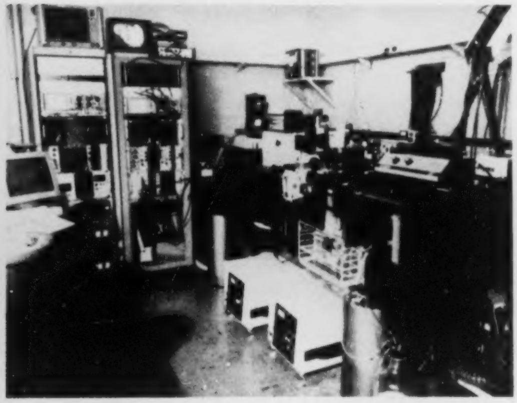


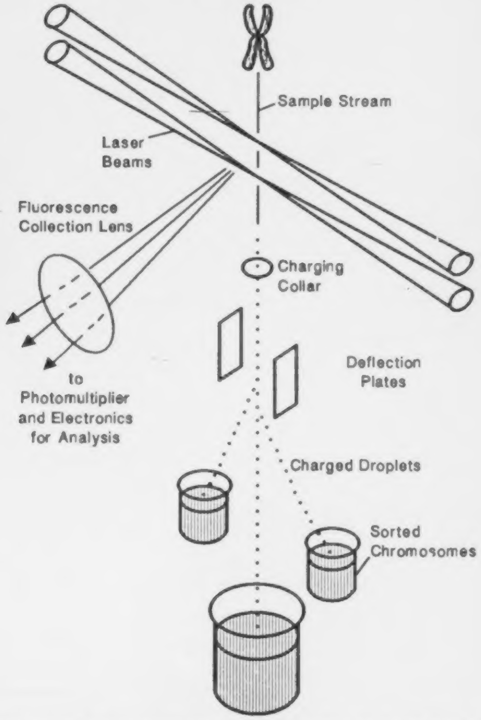
Photo credit: Larry Deaven, Los Alamos National Laboratory, Los Alamos, NM

Flow cytometry facility for chromosome sorting at Los Alamos National Laboratory.

cytometry and purified by flow sorting. Fluorescent markers that bind to chromosomes are used in flow cytometry as the basis of separating chromosomes from one another in a flow sorter (figure 2-8) (21,29,30,46). Because human chromosomes differ in the degree to which they bind the fluorescent markers, it is possible to use this approach to physically separate some chromosomes from others. The dual-laser chromosome sorter has been used successfully to separate all the human chromosomes except chromosomes 10 and 11. In addition, chromosomes from cell lines with translocations and deletions can be used to narrow the location of the gene to a certain chromosomal region (46).

To determine on which chromosome a gene lies, chromosomes are sorted onto different paper filters made of nitrocellulose. There the DNA is denatured and hybridized with a radioactively labeled DNA probe complementary to the gene to be mapped. (In general, the cDNA is available for use as a probe for the gene.) On whichever chromosome the gene appears, the two sequences will hybridize, and the hybridization can be observed using autoradiography.

Figure 2-8.—Chromosome Purification by the Flow Sorter



Chromosomes stained with a fluorescent dye are passed through a laser beam. Each time, the amount of fluorescence is measured and the chromosome deflected accordingly. The chromosomes are then collected as droplets.

SOURCE: Courtesy of Los Alamos National Laboratory, Los Alamos, NM.

Karyotyping

At a stage of cell division when chromosomes have duplicated but not yet separated from one another, they condense to form structures with features that can be observed under a light microscope. The structure of human chromosomes can be studied by chemically fixing white blood cells at the appropriate stage of cell division and then photographing the chromosome spreads as they appear on slides under the microscope. Individual chromosomes are identified in the photograph, cut out, and, in the case of autosomes, matched with their morphologically identical chromosome partner to generate a *karyotype*. Karyotyping has been most useful for correlating gross chromosomal abnormalities with the characteristics of specific diseases (e.g., Down's syndrome and Turner's syndrome).

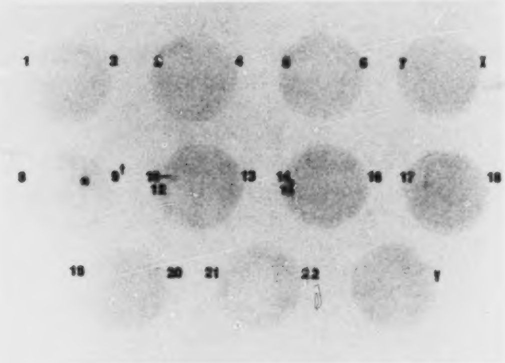


Photo credit: Roger Lebo, the University of California, San Francisco Medical Center. Reprinted with permission from Alan R. Liss, Inc.

Assignment of a gene and genes with related sequences to specific human chromosomes. Samples of the 21 different human chromosomes were sorted onto 11 circular filters and then hybridized to a radioactively labeled DNA probe from the aldolase gene (aldolase is an enzyme involved in the metabolism of sugars). Most of the radioactive signal in the autoradiograph appears on the filter with chromosome 9, indicating that the complementary DNA sequence is in that chromosome. The autoradiograph also shows some hybridization to chromosomes 17 and 10, indicating that aldolase genes with different, but similar, sequences are found on these chromosomes.

Chromosome Banding

Using fluorescent dyes as chromosome-specific stains, Caspersson and others (10-12) developed optical methods for observing the banding patterns on human chromosomes. These methods reveal more details of chromosome morphology than does simple karyotyping. The bands are chromosomal regions that appear as stripes on chromosome spreads when viewed under the light microscope. Each of the 24 different human chromosomes has a unique banding pattern, thus the bands can be used to identify individual chromosomes. Genes can be mapped to specific bands by identifying differences between the banding patterns on chromosomes from normal individuals and those on the chromosomes from an individual with a significant chromosomal alteration.

Nearly 1,000 distinct bands have been detected on the 24 human chromosomes by staining and light microscopy, and an average of 100 genes is represented in a single band (50). Chromosome banding is a useful procedure for finding the general location of a gene, but it does not offer sufficient resolution to identify the exact position of a gene relative to other genes mapped in the same region (58).

In Situ *Hybridization*

Family linkage and somatic cell hybridization are not direct mapping methods; they are based on the correlation between traits and the frequency of transmission of those traits in families. Karyotyping and analysis of chromosome banding allow a specific trait to be correlated with a particular chromosome or a large region of a chromosome. Advances in molecular biology have overcome the limitations of those techniques by providing means for more precise mapping of genetic markers. One such method is *in situ* hybridization of isolated genes or gene fragments to chromosomal DNA.

The *in situ* hybridization technique was originally developed by Mary Lou Pardue and Joseph Gall for detection of genes encoding ribosomal RNAs in chromosomes from *Drosophila* salivary glands (56). In the typical *in situ* hybridization experiment, the DNA corresponding to a particular gene or gene fragment is used to probe for complementary sequences in chromosomes (28). The chromosomes to be analyzed are fixed on a microscope slide, where the DNA strands are chemically treated and separated. Next, the radioactively labeled DNA probe is mixed with the chromosomes on the slide. Under proper conditions, the DNA probe hybridizes with the gene sequence wherever it is located on the prepared chromosomes.

Results of *in situ* hybridization can be seen by exposing the slides to a photographic emulsion for a long period, then analyzing the photographs under a microscope. Wherever the radioactively labeled DNA strands have paired with complementary chromosomal regions, tiny silver grains appear. The location of a specific gene can be found by counting the number of grains in each region and using computer methods to analyze the data (58). Although *in situ* hybridization has been a principal method for the mapping of human genes to autosomes, higher-resolution methods are nec-

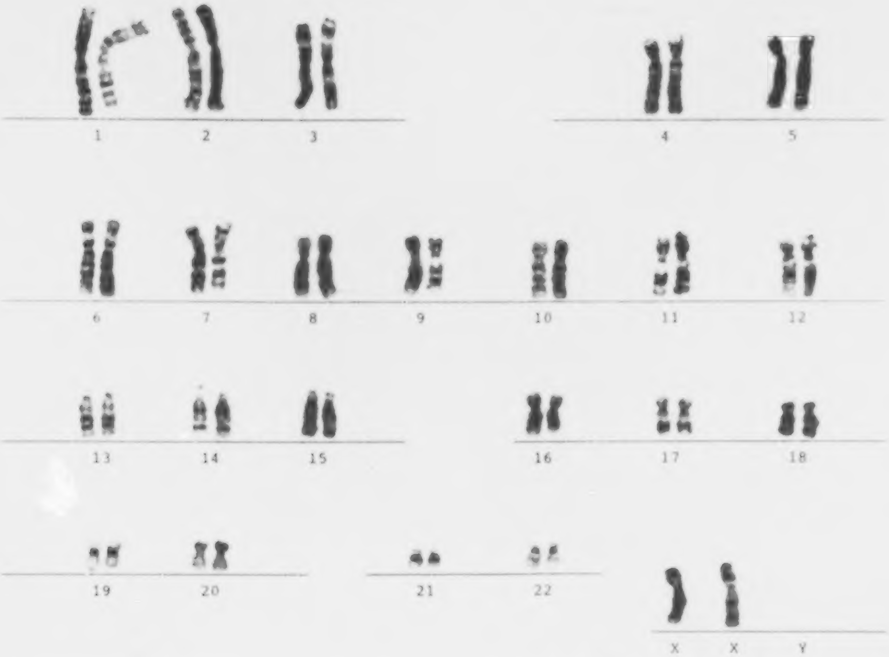


Photo credit: The Genetics and IVF Institute, Fairfax, VA

Human karyotypes of a normal female.

essary. The procedure is limited to a resolution of about 10 million base pairs, a substantial portion of the total length of most chromosomes. Since many genes could fit into such a region, the exact location of the gene of interest must still be determined precisely (58).

Other Methods for Mapping Genes

Several other techniques for mapping human genes are available, including gene dosage mapping and comparative mapping of species. In gene dosage mapping, a correlation is made between the amount of gene product and the presence of extra genes or the absence of a gene or chromosome fragment. Biochemical analysis of cellular contents isolated from an individual with a particular genetic disease, or from somatic cell hybrid lines derived from that individual's cells, is performed to measure amounts of gene products. The structure of the altered chromosome (or chromosomes) is then characterized by one or more of the methods already described.

Comparative mapping of species can provide useful human gene mapping information. This is particularly true among mammals, where it has been demonstrated that different species have similar patterns of gene organization on certain chromosomes [Computer Horizons, Inc., see app. A]. For example, tabulations show that all of the human autosomes except chromosome 13 have at least two linked genes which are also linked in the mouse (35).

Comparison of the banding patterns of chromosomes from different species have also proved useful in matching chromosomes between species, even though differences in total numbers of chromosomes exist. There is, for example, a striking resemblance between chimpanzee and human

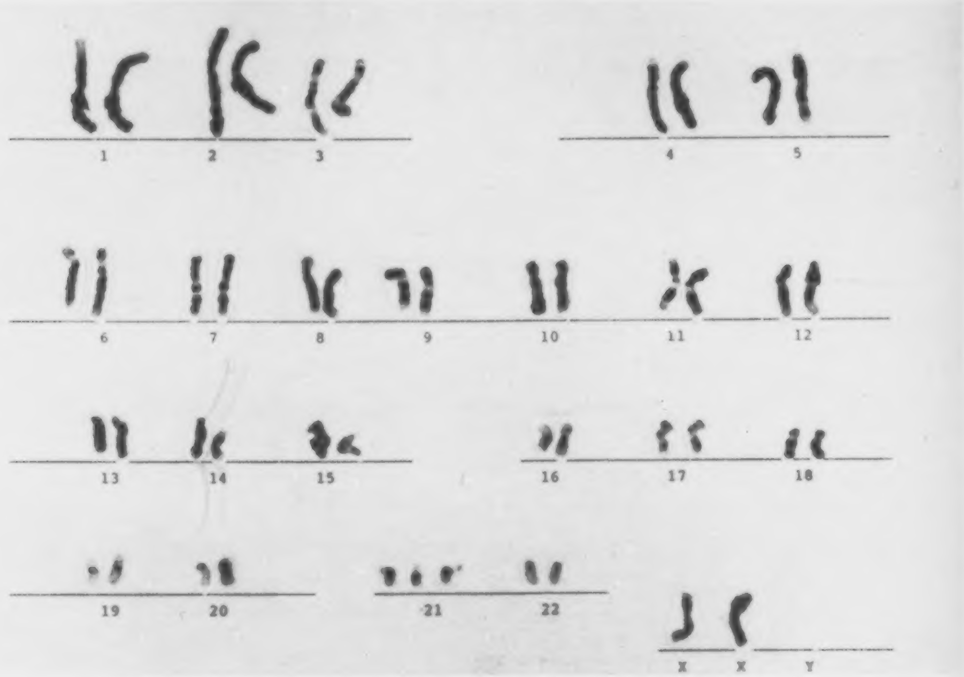


Photo credit: The Genetics and IVF Institute, Fairfax, VA

A female with the extra chromosome 21 associated with Down's syndrome.

chromosomes (81). In recent years, DNA sequence comparisons between widely differing organisms have been used to isolate or confirm the identity of specific human genes (see ch. 3).

HIGH-RESOLUTION PHYSICAL MAPPING TECHNOLOGIES

Construction of high-resolution physical maps of whole genomes involves cutting the component DNA with restriction enzymes, analyzing the chemical characteristics of each fragment, and then reconstructing the original order of the fragments in the genome. Generally, the DNA fragments to be ordered are isolated from chromosomes; united with carrier, or *vector*, DNA molecules originating from viruses, bacteria, or the cells of higher organisms; and introduced into suitable host cells, where the isolated DNA can be reproduced in large quantities. A fragment of DNA is said to be *cloned* when it is stably maintained as part of a DNA vector in a single line of cells. A set of clones representing overlapping segments of DNA encompassing an entire genome is called a *genomic library*. In order to make a physical map, the clones in the genomic library must be ordered in relation to one another's position on the chromosome. The following sections describe in more detail the methods currently available for creating high-resolution physical maps and their application to the genomes of specific organisms.

Cloning Vectors as Mapping Tools

Any genome mapping project first requires the isolation, usually by cloning technologies, of fragments of chromosomal DNA. Several different

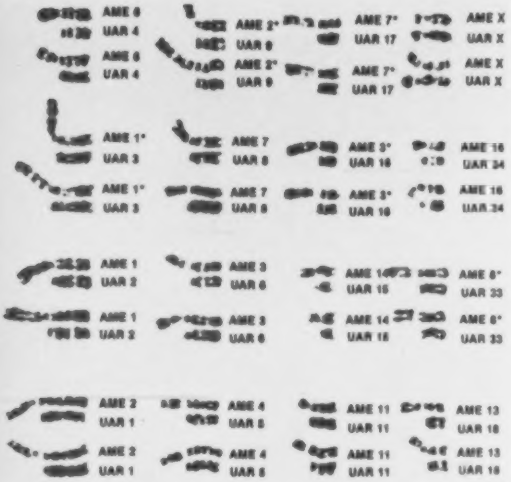

AME 6
UAR 4
AME 6
UAR 4
AME 2*
UAR 9
AME 2*
UAR 9
AME 7*
UAR 17
AME 7*
UAR 17
AME X
UAR X
AME X
UAR X
AME 1*
UAR 3
AME 1*
UAR 3
AME 7
UAR 5
AME 7
UAR 5
AME 3*
UAR 16
AME 3*
UAR 16
AME 16
UAR 34
AME 16
UAR 34
AME 1
UAR 2
AME 1
UAR 2
AME 3
UAR 6
AME 3
UAR 6
AME 14
UAR 15
AME 14
UAR 15
AME 8*
UAR 33
AME 8*
UAR 33
AME 2
UAR 1
AME 2
UAR 1
AME 4
UAR 5
AME 4
UAR 5
AME 11
UAR 11
AME 11
UAR 11
AME 13
UAR 16
AME 13
UAR 16

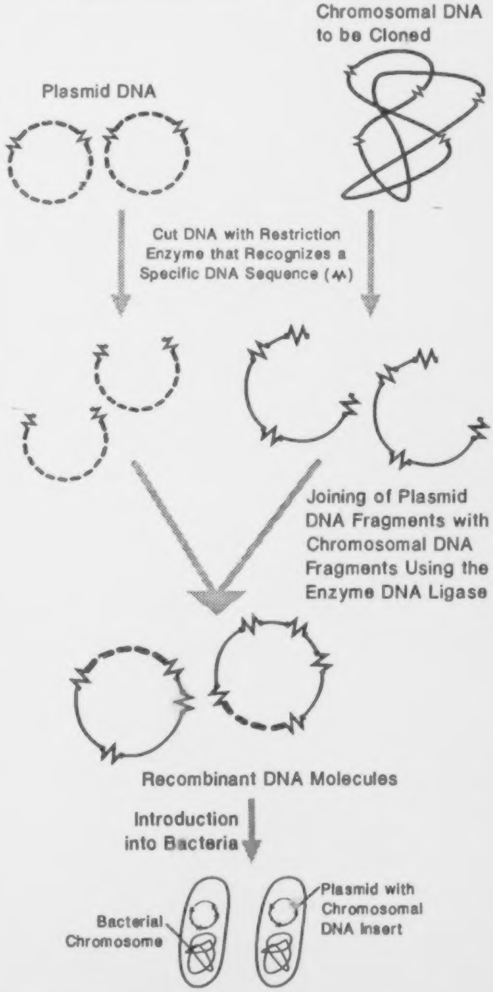
Photo credit: Stephen O'Brien, The National Cancer Institute, Frederick, MD. Reprinted with permission from Nature *317:140-144, 1985.*

A comparative alignment of chromosomes from the giant panda (AME) and the brown bear (UAR). The putative matches between the whole chromosomes, or chromosome segments, of each animal were based on the thickness of stained bands on the chromosomes and on the spacing and intensity of the bands. This type of molecular information has been used to establish the phylogeny of these bears and to demonstrate some of the problems in using the appearance of animals, instead of their chromosome structure, in studies of evolution.

types of cloning vectors have been developed using recombinant DNA technology:

- *Plasmid vectors* are circular DNA molecules of 1,000 to 10,000 base pairs that can carry additional DNA sequences in fragment inserts up to 12,000 base pairs (2,4). Plasmids exist as minichromosomes in bacterial cells (usually between 10 to 100 copies per cell) and are separate from the main bacterial chromosome.
- *Phage lambda chromosomes* are about 50,000 base pairs and can accept foreign DNA inserts up to about 23,000 base pairs (33,79). Just as viruses infect human cells, phage infect bacterial cells and generate hundreds of descendants.
- *Cosmid vectors* are plasmids that also contain specific sequences from the bacterial phage lambda. Cosmids are about 5,000 base pairs, but because they contain phage lambda sequences, they can carry DNA inserts up to about 45,000 base pairs (figure 2-9) (25,34,36).
- *Yeast artificial chromosomes* are plasmids containing portions of yeast chromosomal DNA that function in replication. These artificial chromosomes can accommodate foreign DNA fragment inserts nearly 1 million base pairs long (6).

Figure 2-9.—DNA Cloning in Plasmids


Chromosomal DNA to be Cloned
Plasmid DNA
Cut DNA with Restriction Enzyme that Recognizes a Specific DNA Sequence (ʌʌ)
Joining of Plasmid DNA Fragments with Chromosomal DNA Fragments Using the Enzyme DNA Ligase
Recombinant DNA Molecules
Introduction into Bacteria
Bacterial Chromosome
Plasmid with Chromosomal DNA Insert

SOURCE: Office of Technology Assessment, 1988.

Most of the physical mapping work carried out to date has employed bacteriophage and cosmid cloning vectors because the yeast artificial chromosome vectors have only recently been developed [Myers, see app. A].

Physical Mapping of Restriction Enzyme Sites

With the exception of DNA sequencing, restriction enzyme mapping is the method that gives the highest-resolution picture of DNA as it is organized in a chromosome. Several basic steps are involved in the construction of this type of physical map for part or all of a genome:

- purifying chromosomal DNA,
- fragmenting DNA by restriction enzymes,
- inserting all the resulting DNA fragments into DNA vectors to establish a collection (library) of cloned fragments, and
- ordering the clones to reflect the original order of the DNA fragments on the chromosome.

Variations in any of these steps can affect the resolution of the physical map.

Purification of Chromosomal DNA

Whole chromosomes are the best source of DNA for genomic libraries. Mixtures of chromosomes can be extracted directly from cells, but for organisms with complex genomes, such as human beings, it might be desirable to first separate the different chromosomes and then create sets of clones from the individually purified chromosomes.

Mixtures of whole chromosomes extracted from human cells can be sorted by flow cytometry. Somatic cell hybrid lines carrying one or a few human chromosomes can also be used as a highly enriched source of particular chromosomes. **The refinement of existing methods and the development of new technologies for obtaining large amounts of purified human chromosomes will be crucial in the early stages of human genome mapping projects.**

Fragmentation of DNA

The availability of chromosome fragments of decreasing size allows mapping at higher resolution. A technology called pulsed field gel electrophoresis (PFGE) allows separation of DNA molecules ranging in size from 20,000 to 10 million or more base pairs (8,9,13,61) [Myers, see app. A].

During PFGE, large DNA fragments are subjected to an electric field that is switched back and forth across opposite directions for short pulses of time. This alteration in the direction of the electric field allows very large DNA molecules (up to tens of millions of base pairs) to migrate into the agarose gel and separate from one another, even though the normal size limit for electrophoretic separation of DNA molecules during conventional agarose gel electrophoresis is about 50,000 base pairs. This method is so powerful that it has been used successfully to separate all 14 of the yeast chromosomes from each other (figure 2-10) [Myers, see app. A]. Since intact human chromosomes have an average size of approximately 100 million base pairs, the PFGE technique is only useful for separating large fragments made from individual, purified human chromosomes.

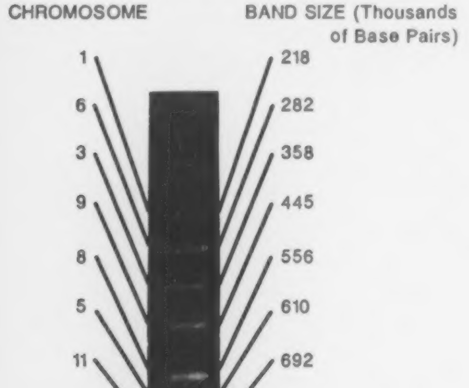
The level of detail possible on a physical map depends on the restriction enzyme or enzymes used. There are a few restriction enzymes that cut DNA very infrequently, generating small numbers of large fragments (ranging from several thousand to a million base pairs). Most restriction enzymes cut DNA more frequently, generating large numbers of small fragments (ranging from fewer than a hundred to greater than a thousand base pairs). The relative order of a small set of large fragments is easier to determine than the order of a large set of short fragments, but it gives a lower-resolution physical map. The choice of enzyme thus depends on the purpose of the physical map. If the aim is to have fragments of a size amenable to DNA sequencing, then a mapped restriction site at least every 500 base pairs would be ideal, but a mapped site every 2,000 to 3,000 base pairs would also be practical. **Given the technology currently available, sequencing the DNA of the 3-billion-base-pair haploid human genome might require the prior mapping**

of as many as 6 million restriction enzyme cutting sites (69) [Myers, see app. A].

Construction of Libraries of DNA Fragments

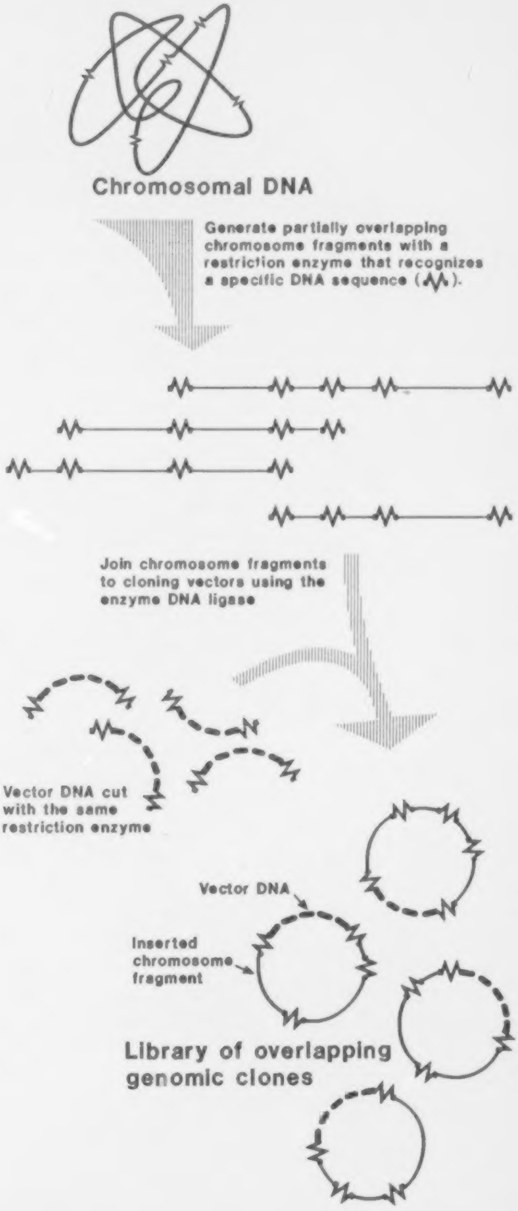
For physical mapping projects, it is important to have as much DNA as needed. The use of cloned DNA fragments offers this advantage. Fragments of DNA from whole chromosomes are generally cloned into vectors such as plasmids, cosmids,

Figure 2-10.—Separation of Intact Yeast Chromosomes by Pulsed Field Gel Electrophoresis


CHROMOSOME
BAND SIZE (Thousands of Base Pairs)
1
218
6
282
3
358
9
445
8
556
5
610
11
692
10
761
14
800
2
834
13
950
16
970
7
1125
15
1125
4
1600
12
2500
Sample Loading Well

SOURCES: Chris Traver and Ronald Davis, California Institute of Technology, Pasadena, CA.

Figure 2-11.—Constructing a Library of Clones Containing Overlapping Chromosomal Fragments


Chromosomal DNA
Generate partially overlapping chromosome fragments with a restriction enzyme that recognizes a specific DNA sequence (ʌʌ).
Join chromosome fragments to cloning vectors using the enzyme DNA ligase
Vector DNA cut with the same restriction enzyme
Vector DNA
Inserted chromosome fragment
Library of overlapping genomic clones

SOURCE: Office of Technology Assessment, 1988.

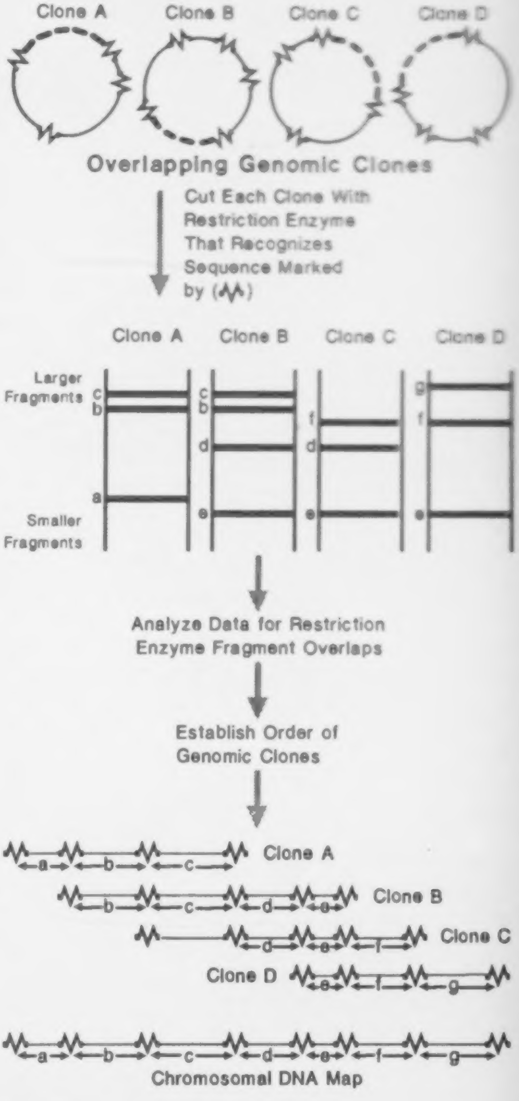
phage chromosomes, and artificial yeast chromosomes. These vectors can be stably maintained in host cells (bacteria or yeast) that multiply rapidly to provide the amounts of DNA necessary for restriction enzyme mapping and DNA sequencing. DNA fragments are usually cloned by cutting the vector of choice with a restriction enzyme and then connecting the newly generated ends of the vector to the ends of the DNA fragments with the enzyme DNA ligase. The resulting collection of clones is called a *library*. There is no obvious order to the library, and the relationship between the components can only be established by physical mapping.

In order to establish that any two clones represent chromosomal segments that normally occur next to one another in the genome, it is necessary to have collections of clones representing partially overlapping regions of chromosomal DNA (figure 2-11). To create libraries of overlapping clones, the chromosomal DNA is treated with a frequent-cutting restriction enzyme, one that cuts every 500 base pairs or so, but conditions are controlled so that the enzyme is not allowed to cut the DNA at all the possible restriction enzyme sites. Instead, by lowering the amount of restriction enzyme used, only partial cutting is allowed. The experimental conditions for partial cutting are adjusted so that DNA fragments are generated with an average size equal to the vector's capacity (usually 20,000 to 50,000 base pairs). In theory, no one of the cutting sites will be recognized by the restriction enzyme more frequently than another, so a population of overlapping segments representing all possible cutting sites in the original DNA sample should be generated. These fragments are then cloned in the appropriate vector.

Determination of the Order of Clones

The clones in a library are ordered by subdividing the chromosomal DNA inserts into even smaller fragments and identifying which clones have some common subfragments. Figure 2-12 illustrates how this is done. A particular DNA clone (vector plus the chromosomal DNA insert) is cleaved with one or more restriction enzymes (other than that used to make the clones) under conditions in which all sites are recognized and cut. The resulting fragments are then run on a gel made of agarose. After electrophoresis, a pattern of fragments is observed along the length of the gel. If the DNA fragments are present in sufficient amounts, they can be seen under ultraviolet light after staining the gel with the dye

Figure 2-12.—Making a Contig Map


Clone A
Clone B
Clone C
Clone D
Overlapping Genomic Clones
Cut Each Clone With Restriction Enzyme That Recognizes Sequence Marked by (ʌʌ)
Clone A
Clone B
Clone C
Clone D
Larger Fragments
Smaller Fragments
c
b
a
c
b
d
e
f
d
e
g
f
e
Analyze Data for Restriction Enzyme Fragment Overlaps
Establish Order of Genomic Clones
a
b
c
Clone A
b
c
d
e
Clone B
d
e
f
Clone C
Clone D
e
f
g
a
b
c
d
e
f
g
Chromosomal DNA Map

SOURCE: Office of Technology Assessment, 1988.

ethidium bromide; otherwise, the phosphates at the ends of the DNA fragments are labeled with a radioactive isotope and viewed after autoradiography. A unique pattern of bands appears (corresponding to DNA fragments) for any given clone because of the unique arrangement of restriction enzyme sites in the region of the chromosome from which that clone was derived. If two clones contain overlapping segments of DNA, then a portion of the banding pattern for each will be identical. For example, if the clone order is A-B-C-D, then the restriction enzyme fragments from clone A will partially overlap with those from clone B, clone B fragments with clone C, and so on (figure 2-12).

Groupings of clones representing overlapping, or contiguous, regions of the genome are known as *contigs* (18,66). On an incomplete physical map, contigs are separated by gaps where not enough clones have been mapped to allow the connection of neighboring contigs. Of all the steps in physical mapping, the connection of all the contigs is the one that faces the greatest number of technical problems. Therefore, **the time required to achieve a complete physical map of any ge-**

GENOMIC MAP OF BACTERIOPHAGE T4

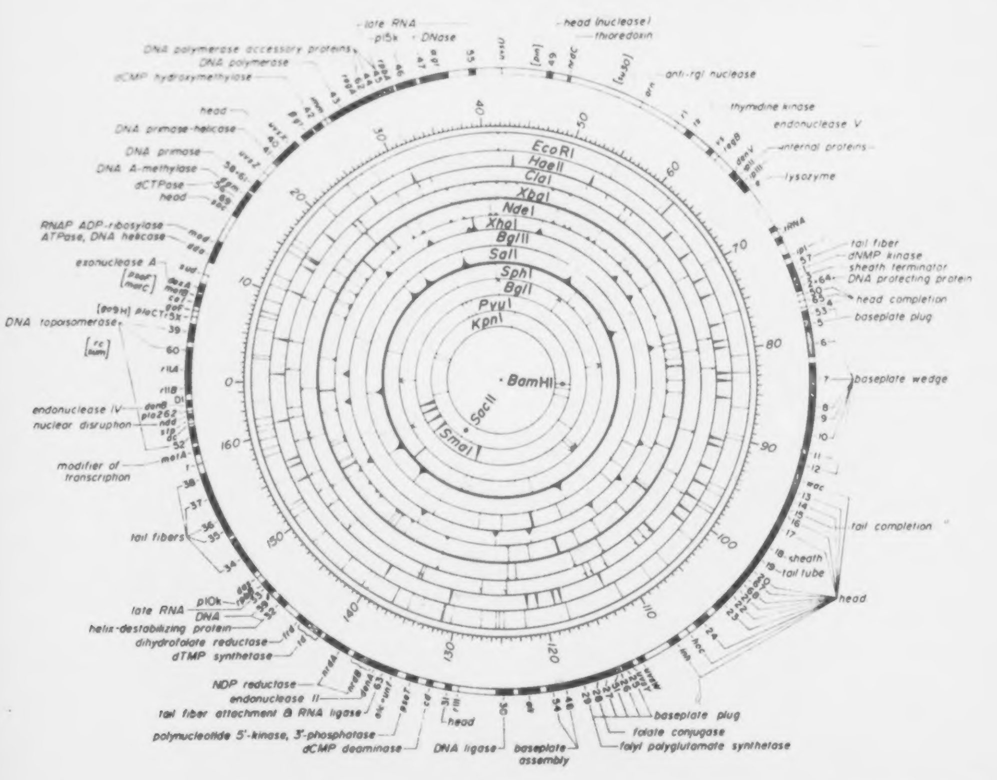

late RNA
p15k
DNase
DNA polymerase accessory proteins
DNA polymerase
dCMP hydroxymethylase
head
DNA primase-helicase
DNA primase
DNA A-methylase
dCTPase
head
RNAP ADP-ribosylase
ATPase, DNA helicase
exonuclease A
DNA topoisomerase
endonuclease IV
nuclear disruption
modifier of transcription
tail fibers
late RNA
p10k
DNA
helix-destabilizing protein
dihydrofolate reductase
dTMP synthetase
NDP reductase
endonuclease II
tail fiber attachment & RNA ligase
polynucleotide 5'-kinase, 3'-phosphatase
dCMP deaminase
head
DNA ligase
baseplate assembly
baseplate plug
folate conjugase
folyl polyglutamate synthetase
head (nuclease)
thioredoxin
anti-rgl nuclease
thymidine kinase
endonuclease V
internal proteins
lysozyme
tRNA
tail fiber
dNMP kinase
sheath terminator
DNA protecting protein
head completion
baseplate plug
baseplate wedge
tail completion
sheath
tail tube
head
EcoRI
HaeII
ClaI
XbaI
NdeI
XhoI
BglII
SalI
SphI
BglI
PvuI
KpnI
BamHI
SacII
SmaI
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160

Photo credit: Elizabeth Kutter and Burton Guttman, Evergreen State College, Olympia, WA. Reprinted with permission from Stephen J. O'Brien (ed.), Genetic Maps 1987, vol. 4, (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1987).

Genomic map of bacteriophage T4. The genome of bacteriophage T4 contains about 166,000 base pairs. Shown are maps illustrating the names of mapped genes (outer circle), genetic map distances (second largest circle), which are measured in *minutes* in bacteria and their phage, and positions of DNA cutting sites for a variety of restriction enzymes (inner circles).

nome is a function of the time required to connect neighboring contigs.

Physical Mapping of Nonhuman Genomes

So far, high-resolution mapping of entire genomes has focused on nonhuman organisms. Most of the technologies applicable to human genetic and physical mapping, therefore, have been developed from work on other organisms. Mapping of complete genomes is well underway for several species of bacteria and yeast, for the nematode, and is beginning for the fruit fly. These organisms have long served as excellent model systems for what are sometimes found to be universal genetic and biochemical mechanisms governing cell physiology. The technologies employed in these high-resolution genome mapping projects range from making contig maps to fine mapping by DNA sequencing.

The Bacterial Genome

For many years, bacteria (mainly *Escherichia coli*) and phage (viruses that infect bacteria) have been principal research subjects of molecular biologists, molecular geneticists, and biochemists. Because of the relative ease of studying gene function in *E. coli*, it is the organism whose genetics and biochemistry are closest to being completely understood. The DNA of this bacterium is contained in a single circular chromosome of 4.7 million base pairs (63). The genetic map of *E. coli* is quite extensive, with about 1,200 of the 5,000 or so known genes already cloned (3). In addition, the nucleotide sequence of over 20 percent of this bacterial genome is known (26).

Progress on the physical map of the *E. coli* genome is good. Cassandra Smith and co-workers at Columbia University made a complete physical map of this genome using a restriction enzyme called Not I, which cuts DNA only infrequently (63). Not I recognizes a sequence eight nucleotides long that is expected to occur by chance once every 34,000 base pairs. Only 22 Not I sites were found in the *E. coli* genome (63).

A higher-resolution physical map of *E. coli* was generated by Kohara and colleagues (42) at Nagoya University in Japan. These researchers devised an innovative, rapid mass-analysis mapping approach involving eight different restriction enzymes. In a period of time equivalent to only one-half of a person-year, this group generated a high-resolution physical map covering 99 percent of the *E. coli* genome, leaving only seven gaps. An independent effort to generate a high-resolution map of the cutting sites for three different, frequent-cutting restriction enzymes is also near completion at the University of Wisconsin, Madison (20).

The work of the researchers in Wisconsin and Japan is important because it generates an ordered set of clones. A map indicating the order of a library of genomic clones is immediately useful to anyone wishing to examine DNA corresponding to a gene whose position on the map is known. The physical map is correlated with the genetic map at many sites in *E. coli*, primarily as a result of including in the analysis clones containing known genes. Kohara and co-workers demonstrated that the use of large fragments for connecting groups of clones is not necessary for *E. coli*. Because of the computational limitations on connecting great numbers of small fragments, however, large-fragment maps, analogous to the Not I map of *E. coli*, will no doubt play a significant role in mapping large genomes, such as the human genome.

The Yeast Genome

An ongoing project to map the 15 million base pairs in the *Saccharomyces cerevisiae* (baker's yeast) genome has been described by Olson and colleagues (55) at Washington University. These researchers initiated the mapping project to facilitate the organization of the vast amount of information already available on this organism. As Olson writes:

> Just as conventional cartography provides an indispensable framework for organizing data in fields as diverse as demography and geophysics, it is reasonable to suppose that "DNA cartography" will prove equally useful in organizing the vast quantities of molecular genetic data that may be expected to accumulate in the coming decades (55).

A large fraction of the *S. cerevisiae* genome (about 95 percent) is available in clones that have been joined together in over 400 contiguously mapped stretches. These contigs are being correlated with a complete large-fragment restriction map for the yeast genome. These combined maps make it possible to construct or identify a mapped region 30,000 to 100,000 base pairs in length around virtually any starting point, typically a cloned gene [Mount, see app. A].

The Nematode Genome

The nematode *Caenorhabditis elegans* is a popular organism among developmental biologists because the origin and function of all 958 cells in the adult animal are known, offering researchers the opportunity to study the basis of organismal development. With its 3-day generation time, *C. elegans* is also suited to genetic studies. Molecular biologists, interested in the molecular basis of development, would find an ordered set of clones from the nematode genome particularly useful for their work [Mount, see app. A].

Coulson and Sulston at the Medical Research Council in England initiated a *C. elegans* mapping project to provide such tools and to establish communications among the laboratories working on this organism. Like the *S. cerevisiae* genome mapping project, this resulted in a set of clones that covers most of the genome (18). One difference is that the *C. elegans* clones are put into order by the fingerprinting method: Distances from each cleavage site for one enzyme to the nearest site for a second enzyme were measured, and clones sharing a number of such distances (measured as lengths of restriction fragments observed on polyacrylamide gels) were considered to overlap. This process makes identification of overlapping regions somewhat easier (because the information is denser), at the expense of more precise physical map information. A second difference is that cosmid clones were used in the nematode project, while phage clones were used in the yeast project. Cosmid clones can accommodate larger DNA inserts than phage clones, but they can also be less stable, with portions of the inserts becoming deleted more often (17). At present, over 700 contigs, ranging from 35,000 to 350,000 base pairs in length and representing 90 percent of the *C. elegans* genome, have been characterized (71).

The Fruit Fly Genome

The genetics of the common fruit fly, *Drosophila melanogaster*, are the best characterized of any multicellular organism. One reason for studying fruit flies is that it is possible to carry out a saturating screen to detect mutations of a particular type. In a saturating screen, every gene that could mutate to produce the defect being studied is identified. (This accomplishment is crucial to a complete understanding of many cellular processes.) The saturating screen technique allows for a comprehensive genetic analysis because the entire genome can be examined for the presence of genes that are involved in a particular process. The most celebrated example is an exhaustive study of mutations that are lethal to the fly in its larval stage (39,53,78) [Mount, see app. A].

Until recently, the physical mapping of the 165 million base pairs in the *D. melanogaster* genome had not been undertaken by any one laboratory. Roughly 500 to 1,000 genomic clones have been obtained in various laboratories in various vectors; all of these clones have been localized to a chromosomal map position by *in situ* hybridization to polytene chromosomes (a multicopy set of *D. melanogaster* chromosomes unique to its salivary gland). A listing of these clones is maintained by John Merriam and colleagues at the University of California, Los Angeles, and the clones are made available to all researchers [Mount, see app. A].

Work by Michael Ashburner and co-investigators at Cambridge University on a comprehensive map of overlapping cosmid clones of the *D. melanogaster* genome was approved for funding by the European Economic Community in late 1987. This project is expected to follow the fingerprinting strategy of the nematode project, with the important difference that cytological maps (maps of banding patterns derived from microscopic analysis of stained chromosomes) of *D. melanogaster* chromosomes will be exploited. First, the technique of microdissection cloning (whereby DNA is excised from precise regions of the salivary gland polytene chromosomes and cloned) will be used to generate region-specific genomic clones. These microdissection clones are not of sufficient quality to be used directly, but they can be used to correlate cosmids in a standard genomic library with specific chromosomal regions. This step makes it easier to assemble the contiguous clones into groups. Finally, the position of all contigs with respect to the cytological map will be confirmed by *in situ* hybridization, whereby cosmid clones from the various contigs would be hybridized to salivary gland chromosomes [Mount, see app. A].

Strategies for Physical Mapping of the Human Genome

It is likely that making contig maps of large genomes, such as the human genome, will require a combination of bottom-up mapping and top-down mapping (55). Bottom-up mapping starts by making genomic clones, then fragmenting these clones further to decipher the overlaps necessary for connecting clones into contigs. Top-down mapping (e.g., Smith's *E. coli* map) is of lower resolution because it is derived from minimal fragmentation of source DNA. The critical distinction between the two methods is the size of the genomic DNA fragments used. Bottom-up mapping starts with relatively small genomic clones, while top-down mapping starts with large genomic clones. The advantage of top-down mapping is that it offers more continuity (fewer gaps), while the bottom-up method has higher resolution (more detail). In formulating strategies for mapping the human genome, it will be necessary to decide what level of molecular detail is necessary to begin a human genome mapping project. Will information-rich strategies like those used to develop high-resolution *E. coli* restriction enzyme maps (20,42) or the DNA signposts offered by a RFLP map be the best first-generation human genome maps?

Contig Mapping

Scientists in the fields of molecular biology and human genetics who reviewed an OTA contract report on possible strategies for making contig maps of the human genome [Myers, see app. A] favored the following strategy: to map the genome one chromosome at a time, dividing and subdividing each chromosome into smaller and smaller segments before beginning restriction enzyme mapping and ordering of clones. After subdivision, restriction maps of these smaller segments would be determined and the information linked together to form continuous maps of whole chromosomes. In principle, this strategy could be broken down into five consecutive steps:

1. isolation of each human chromosome,
2. division of each chromosome into a collection of overlapping DNA fragments 0.5 to 5 million base pairs in length,
3. subdivision and isolation of each of these chromosomal fragments into overlapping DNA fragments about 40,000 base pairs in length,
4. determination of the order of the 40,000-base-pair DNA fragments as they appear in the chromosomes and determination of the positions of cutting sites for a restriction enzyme within each of these fragments, and
5. use of the mapping information gained in step 4 to link together each of the overlapping 0.5- to 5-million-base-pair fragments isolated in step 2 [Myers, see app. A].

The substantial progress made so far on contig maps of nonhuman genomes implies that technologies already exist to begin construction of a global physical map of the human genome. The haploid human genome (approximately 3 billion base pairs) is at least 30 times larger than that of the nematode, the largest genome for which comprehensive physical mapping has been attempted. Sulston predicted that the mapping work he and his co-workers have done over the past 4 years could be repeated within 2 person-years, because much of their time was spent devising computer methods for data analysis (17). If the size of a genome were linearly related to the time required to physically map it, then the human genome could be mapped to the same degree of completion as the nematode genome (90 percent) in about 60 person-years. Such calculations are simplistic, however, because features of the human genome other than its size make it potentially more difficult to map. For example, some DNA sequences are repeated frequently throughout the human genome, in contrast to the nematode genome, and these are likely to interfere with the physical mapping process.

Techniques for isolating large chromosomal fragments should offer solutions to some of the physical mapping problems expected to arise from the occurrence of repetitive sequences in the human genome. The two most promising methods developed to date are the PFGE technology (8,9,13,61) and the yeast artificial chromosome cloning technology (6).

A National Research Council advisory panel on mapping and sequencing the human genome recommended improvements in technologies for the following to facilitate the construction of physical maps of large genomes:

- separating intact human chromosomes;
- separating and immortalizing identified fragments of human chromosomes;
- cloning the cDNAs representing expressed genes, especially those that represent rare cell-, tissue-, and development-specific mRNAs;
- cloning very large DNA fragments;
- purifying very large DNA fragments, including higher-resolution methods for separating such fragments;
- ordering the adjacent DNA fragments in a DNA clone collection; and
- automating the various steps in DNA mapping, including DNA purification and hybridization analysis, and developing novel methods that allow simultaneous handling of many DNA samples (52).

DNA Sequencing

Strategies for sequencing the entire human genome are much more controversial than those for generating contig maps. Some scientists favor sequencing only expressed genes, identified with a cDNA map (17). Others propose that sequencing should continue to be targeted at specific regions of interest, as is currently done. Still others hold the view that the whole genome should be sequenced because it could reveal sequences with important functions that would otherwise go unidentified (see ch. 3). The National Research Council panel proposed first that pilot programs be conducted with a goal of sequencing approximately 1 million continuous nucleotides (which is about five times as large as the largest continuous stretch of DNA sequenced to date) (52). Second, improvements in existing DNA sequencing technologies would be vigorously encouraged. Finally, extensive sequencing of other genomes, including the mouse, fruit fly, nematode, yeast, and bacterial genomes, was recommended for purposes of comparison (52).

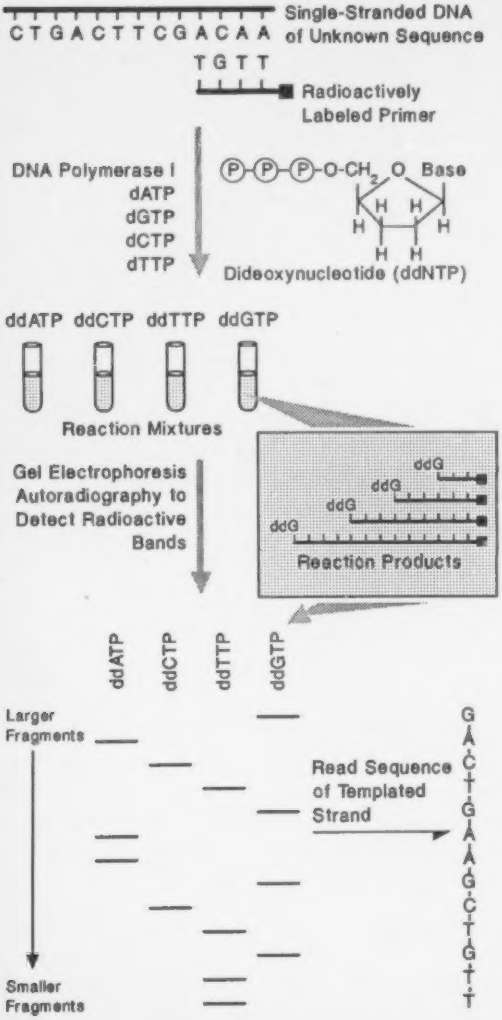
The potential uses of human genome maps and sequences will likely dominate strategic decisions on which of the possible methods should be used to construct them (ch. 3). The strategy currently favored—preparing physical maps of individual chromosomes—requires that decisions be made on which chromosomes should be mapped first. Mapping smaller chromosomes first in pilot projects (e.g., chromosomes 21 and 22) would be the logical strategy from a technical perspective. Alternatively, selecting chromosomes linked to the largest numbers of markers for human genetic diseases (e.g., chromosome 7 and the X chromosome) might make the impact of genome mapping on clinical medicine more immediate. Efforts are already underway in a number of U.S. and foreign laboratories (ch. 8) to physically map (at various levels of resolution) human chromosomes known to be of general clinical significance or to carry genes of specific interest to the researchers involved. Scientists at Los Alamos and Livermore National Laboratories have begun mapping chromosomes 16 and 19, respectively. These chromosomes were chosen for their relatively small sizes and number of clinically relevant genetic markers. Researchers at Columbia University have begun work on a physical map of chromosome 21 for similar reasons.

DNA Sequencing Technologies

Two methods for sequencing DNA are standard in laboratories today. One technique, developed by Fred Sanger and Alan Coulson at the Medical Research Council in England (60), uses enzymes (figure 2-13), while the other, developed by Alan Maxam and Walter Gilbert at Harvard University, involves chemicals that degrade DNA (figure 2-14) (48,49). The two methods differ in the means by which the DNA fragments are produced; they are similar in that sets of radioactively labeled DNA fragments, all with a common origin but terminating in a different nucleotide, are produced in the DNA sequencing reactions.

George Church at Harvard Biological Laboratories has adapted the Maxam and Gilbert DNA sequencing method in an innovative technology,

Figure 2-13.—DNA Sequencing by the Sanger Method


Single-Stranded DNA of Unknown Sequence
C T G A C T T C G A C A A
T G T T
Radioactively Labeled Primer
DNA Polymerase I
dATP
dGTP
dCTP
dTTP
P P P O-CH2 O Base
H H H H H H
Dideoxynucleotide (ddNTP)
ddATP ddCTP ddTTP ddGTP
Reaction Mixtures
Gel Electrophoresis Autoradiography to Detect Radioactive Bands
ddG
ddG
ddG
ddG
Reaction Products
ddATP
ddCTP
ddTTP
ddGTP
Larger Fragments
Smaller Fragments
Read Sequence of Templated Strand
G A C T G A A G C T G T T

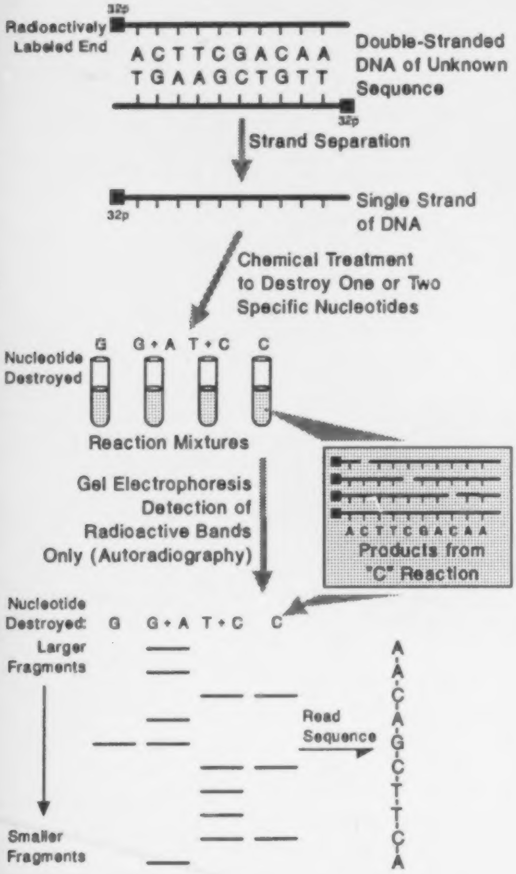
In the Sanger method, a cloned DNA fragment is mixed with a short piece of synthetic DNA complementary to only one end (the origin) of the cloned fragment. An enzyme called DNA polymerase is then used to catalyze the synthesis of a complementary strand. During the polymerization reaction, a modified nucleotide, a dideoxynucleotide, is included with a mixture of the four naturally occurring nucleotides (A, G, T, and C), one of which is labeled with a radioactive phosphorous or sulfur atom, causing growth of the DNA chain to stop whenever the modified nucleotide is inserted. Four separate reactions, each containing all four normal nucleotides but a different dideoxynucleotide, can be carried out. A series of radioactively labeled DNA strands will be made, the lengths of which depend on the distance from the origin to the nucleotide position where the chain was terminated. For example, if a short DNA template has four G's, conditions are set up such that some molecules will be made with no G dideoxynucleotide analogs, some will terminate at the fourth G position, some at the third G position, and so on. Similarly, the other three dideoxynucleotides will insert infrequently and randomly at the appropriate positions in the other three nucleotide-specific reactions. The series of labeled DNA strands is subsequently analyzed by polyacrylamide gel electrophoresis. Radioactively labeled DNA is electrophoresed through a vertical slab of polyacrylamide gel (polyacrylamide is a polymeric resin in which DNA molecules from 1 to 400 bases long can be separated from one another), an X-ray film is then placed over the gel and exposed, and the resulting autoradiograph shows a ladderlike pattern of bands. The sequencing reactions corresponding to each of the four different bases are run as four adjacent lanes on the polyacrylamide gel, and the resulting ladders of bands are read alternately to give the sequence of the DNA.

SOURCE: Office of Technology Assessment, 1988.

called multiplex sequencing, that enables a researcher to analyze a large set of cloned DNA fragments as a mixture throughout most of the DNA sequencing steps. Mixtures of clones are operated on in the same way as a single sample in traditional sequencing. This is accomplished by tagging each DNA clone in the mixture with short, unique sequences of DNA in the first step and then deciphering the nucleotide sequence of each cloned fragment in the final step. Multiplex sequencing allows the simultaneous analysis of about 40 clones on a single DNA sequencing gel, increasing the efficiency of the standard procedure by more than a factor of 10 (14). Church and coworkers have been applying the multiplex sequencing strategy to determine the complete nucleotide sequences of two species of bacteria, *E. coli* and *Salmonella typhimurium* (14).

The major problem with current DNA sequencing technology is the large number of DNA sequences that remains to be determined. Multiplex is only one of several new sequencing protocols that could be of great value to large genome sequencing projects. Church and Gilbert devised a method related to multiplex sequencing that allows sequencing directly from genomic DNA (15). Another method, developed by researchers at Cetus (Emeryville, CA), involves the selective amplification of specific DNA sequences without prior

Figure 2-14.—DNA Sequencing by the Maxam and Gilbert Method


32p
Radioactively Labeled End
A C T T C G A C A A
T G A A G C T G T T
Double-Stranded DNA of Unknown Sequence
32p
Strand Separation
Single Strand of DNA
32p
Chemical Treatment to Destroy One or Two Specific Nucleotides
Nucleotide Destroyed
G
G + A
T + C
C
Reaction Mixtures
Gel Electrophoresis Detection of Radioactive Bands Only (Autoradiography)
A C T T C G A C A A
Products from "C" Reaction
Nucleotide Destroyed:
G
G + A
T + C
C
Larger Fragments
Smaller Fragments
Read Sequence
A A C A G C T T C A

In the Maxam and Gilbert procedure, chemical reactions specific to each of the four bases are used to modify DNA fragments at carefully controlled frequencies. One end of one strand in a double-stranded DNA fragment is radioactively labeled, and the labeled DNA is used in each of four separate reactions and treated with a chemical that specifically nicks one or two of the four bases in the DNA. When these DNA molecules are treated with another chemical, the DNA fragments are broken where the base was nicked and are destroyed. Just as in the Sanger sequencing method, the products of the Maxam and Gilbert sequencing procedure are fragments of varying lengths, each ending at the G, C, T, or A where the chemical reaction took place. By limiting the amount of chemicals used in each of the base-specific reactions so they will react only a few times per molecule, it is possible to obtain all possible double-stranded DNA fragments equal in length to the distance from the radioactively labeled origin to each of the bases. For any given DNA fragment sequenced, each of the four reactions is electrophoresed separately, as described in figure 2-13, and the sequencing patterns determined from the autoradiograph.

SOURCE: Office of Technology Assessment, 1988.

cloning (59). Each of these methods could potentially eliminate the steps of cloning and DNA preparation in sequence analysis (41).

Finally, DNA sequencing methods that do not involve either gel electrophoresis or chemical or enzymatic reactions have also been proposed. At the Los Alamos National Laboratory, researchers are investigating ways to use enhanced fluorescence detection methods in flow cytometry as an alternative to gel techniques for DNA sequencing. Others have suggested scanning tunneling electron microscopes to read bases directly on a strand of DNA (57,62).

AUTOMATION AND ROBOTICS IN MAPPING AND SEQUENCING

The longest single stretch of DNA sequence determined to date, the genome of the Epstein-Barr virus, contains fewer than 200,000 base pairs. The total number of nucleotides sequenced to date using both chemical and enzymatic sequencing technologies is about 16 million base pairs [Computer Horizons, Inc., see app. A]. This is the current size of GenBank®, the U.S. repository of DNA sequence data [app. D]. Since GenBank® includes only reported data, 16 million base pairs represents a low estimate of the total number of base pairs sequenced. Reported DNA sequences range from those of small viruses to those of animals and plants (table 2-3). So far, less than one-tenth of 1 percent (1.9 million base pairs) of the nearly 3 billion base pairs in the haploid human genome has been sequenced and reported (7). The current DNA sequencing rate is estimated to generate only about 2 million base pairs per year of sequence information (7), a powerful incentive for

Table 2-3.—Amount of Genome Sequenced in Several Well-Studied Organisms

Organism	Genome size (base pairs)	Percent sequenced
Escherichia coli (bacterium)	4.7 million	16
Saccharomyces cerevisiae (yeast)	15 million	4
Caenorhabditis elegans (nematode)	80 million	.06
Drosophila melanogaster (fruit fly)	155 million	.26
Mus musculus (mouse)	3 billion	.04
Homo sapiens (human)	2.8 billion	.08

SOURCES:
C. Burks, GenBank®, Los Alamos National Laboratory, Los Alamos, NM, personal communication, March 1988.
GenBank® Release No. 54, December 1987.

devising methods of automating the procedures involved in preparing for and carrying out DNA sequencing. Some recent reviews (16,38,41,47,57) provide detailed accounts of the robotic and automated systems currently available and describe the kinds of systems being developed or planned for genome mapping and sequencing.

Any degree of automation will help lower the overall costs of genome projects, both in time and in dollars. The primary objective in the use of automation is standardization, driven by the need for repetitive, highly accurate determinations (41). Some of the existing automated devices are designed for repetitive DNA cloning steps, such as the preparation and restriction enzyme cutting of cloned DNA samples. Similarly, efforts are being made to automate the pouring, loading, and running of gels for separating DNA and for sequencing DNA. Many of the steps in physical mapping could be adapted to automation. Cloning procedures, DNA probe synthesis, and DNA hybridizations are only a few of those being explored for application to genome projects. A system that automates some steps in growing DNA clones, to be used, for example, as gene probes or for DNA sequencing, was recently introduced by Perkin-Elmer Cetus Instruments (Norwalk, CT) (67).

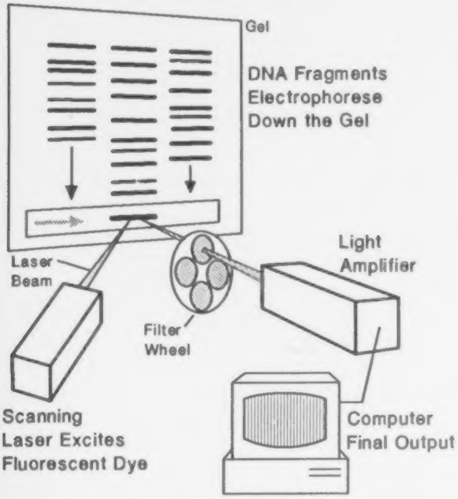
The area of automation that has received the most attention is DNA sequencing. An international workshop on automation of DNA sequencing technologies was held in 1987 in Okayama, Japan, and the proceedings give an extensive account of the state of the art from an international perspective (32). There are five steps in the task of DNA sequence analysis:

1. cloning or otherwise isolating the DNA,
2. preparing the DNA for sequence analysis,
3. performing the chemical (Maxam and Gilbert) or enzymatic (Sanger) sequencing reactions,
4. running the sequencing gels, and
5. reading the DNA sequence from the gel.

Steps 3 through 5 are the functions most often performed by the instruments developed as of early 1988 (16,57,70); however, none of the companies involved has yet commercialized an integrated system that performs all of the functions.

In 1986, Applied Biosystems, Inc. (Foster City, CA) introduced the first commercial, automated DNA sequencer (16). This instrument was made using technology developed by Leroy Hood and co-workers at the California Institute of Technology. This and similar machines perform steps 4 and 5. The Applied Biosystems, Inc. system is based on the Sanger sequencing reaction, with modifications to use different fluorescent dyes instead of radioactive chemicals to label the primers. Because the sequencing reaction primers are individually labeled with different dyes, each of the four enzymatic reactions can be run together in a single lane on the polyacrylamide gel. A laser activates the dyes, and fluorescence detectors read the DNA sequence at the bottom of the gel as each fragment appears. The sequence is determined directly by a computer (figure 2-15). E.I. du Pont de Nemours & Co. (Wilmington, DE) introduced in 1987 an automated system that slightly modifies the technology used by Hood and Applied Biosystems, Inc.; this system can potentially reduce the number of artifacts read by the fluorescence detectors. Hitachi, Ltd. (Tokyo, Japan) is also expected to market an instrument that automates steps 4 and 5, but it too is based on the fluorescence technology developed by Hood and colleagues. In early 1988, another U.S. company, EE&G Biomolecular (Wellesley, MA), began marketing a machine that automates the same DNA sequencing and gel-reading methods used manually in most laboratories. Bio-Rad Laboratories (Richmond, CA) marketed an instrument that

Figure 2-15.—Automated DNA Sequencing Using Fluorescently Labeled DNA


Gel
DNA Fragments Electrophorese Down the Gel
Light Amplifier
Laser Beam
Filter Wheel
Scanning Laser Excites Fluorescent Dye
Computer Final Output

SOURCE: Leroy Hood, California Institute of Technology, Pasadena, CA.

scans autoradiographs of DNA sequencing gels and analyzes the data.

Most of these DNA sequencing systems are based on manual enzymatic sequencing reactions, while the gel running and reading are automated. Only one commercial enterprise, Seiko of Japan, has reported automating the chemical or enzymatic steps (step 3) in the DNA sequencing protocol (70). In addition, the University of Manchester Institute of Science (Manchester, England) has built an automatic reagent manipulating system to carry out the Sanger sequencing reactions (47).

Robotics are used to give automation flexibility, to extend its capabilities to complex operations typically performed by highly skilled laboratory workers. Conceivably, laboratory robots would allow programmable devices to do physical work as well as to process data (41). Several robotic devices have been designed and used successfully by companies involved in the commercialization of recombinant DNA products or processes. Genetics Institute's (Cambridge, MA) Autoprep® Plasmid Isolation System provides small quantities of plasmid DNA and vector DNA for DNA sequencing (22). Researchers at the same company also developed a robot to purify and isolate synthetic oligonucleotides for use as probes in cloning and DNA sequencing (38).

Technical advances are occurring rapidly and simultaneously in biology, robotics, and computer science, so it is difficult to predict what the future will bring in the development of automated technology. Some yet-to-be-developed technology could make many of the current physical mapping procedures obsolete.

CHAPTER 2 REFERENCES

1. Ayala, F.J., "Two Frontiers of Human Biology: What the Sequence Won't Tell Us," *Issues in Science and Technology*, (Spring):51-56, 1987.
2. Bernard, H.U., and Helinski, D.R., "Bacterial Plasmid Cloning Vehicles," in *Genetic Engineering*, vol. 2, J.K. Setlow and A. Hollaender, eds., (New York: Plenum Press, 1980).
3. Blattner, F., University of Wisconsin, Madison, personal communication, October 1987.
4. Bolivar, F., and Backman, K., "Plasmids of *E. Coli* as Cloning Vectors," *Methods in Enzymology* 68:245-267, 1979.
5. Botstein, D., White, R.L., Skolnick, M., et al., "Construction of a Genetic Linkage Map in Man Using Restriction Fragment Length Polymorphisms," *American Journal of Human Genetics* 32:314-331, 1980.
6. Burke, D.T., Carle, G.F., and Olson, M.V., "Cloning of Large Segments of Exogenous DNA Into Yeast Using Artificial-Chromosome Vectors," *Science* 236:806-812, 1987.
7. Burks, C., Los Alamos National Laboratory, Los Alamos, NM, personal communication, February 1988.
8. Carle, G.F., Frank, F., and Olson, M.V., "Electrophoretic Separations of Large DNA Molecules by Periodic Inversion of the Electric Field," *Science* 232:65-68, 1986.
9. Carle, G.F., and Olson, M.V., "An Electrophoretic Karyotype for Yeast," *Proceedings of the National Academy of Sciences USA* 82:3756-3760, 1985.
10. Caspersson, T., Lomakka, C., and Zech, L., "Fluores-

cent Banding," *Hereditas* 67:89-102, 1971.

11. Caspersson, T., Zech, L., and Johansson, C., "Differential Banding of Alkylating Fluorochromes in Human Chromosomes," *Experimental Cell Research* 60:315-319, 1970.
12. Caspersson, T., Zech, L., Johansson, C., et al., "Quinocrine Mustard Fluorescent Banding," *Chromosoma* 30:215-227, 1970.
13. Chu, G., Vollrath, D., and Davis, R.W., "Separation of Large DNA Molecules by Contour-Clamped Homogeneous Electric Fields," *Science* 234:1582-1585, 1986.
14. Church, G.M., "Genome Sequence Comparisons," grant proposal, Feb. 10, 1987.
15. Church, G.M., and Gilbert, W., "Genomic Sequencing," *Proceedings of the National Academy of Sciences USA* 81:1991-1995, 1984.
16. Connell, C., Fung, C., Heiner, J., et al., "Automated DNA Sequence Analysis," *BioTechniques* 5:342-348, 1987.
17. Costs of Human Genome Projects, OTA, workshop, Aug. 7, 1987.
18. Coulson, A., Sulston, J., Brenner, S., et al., "Toward a Physical Map of the Genome of the Nematode *Caenorhabditis elegans*," *Proceedings of the National Academy of Sciences USA* 83:7821-7825, 1986.
19. Crick, F.H.C., and Watson, J.D., "The Complementary Structure of Deoxyribonucleic Acid," *Proceedings of the Royal Society(A)* 223:80-96, 1954.
20. Daniels, D.L., and Blattner, F.R., "Mapping Using Gene Encyclopedias," *Nature* 325:831-832, 1987.
21. Deaven, L.L., Van Dilla, M.A., Bartholdi, M.F., et al., "Construction of Human Chromosome-Specific DNA Libraries Fran. Flow. Sorted Chromosomes," *Cold Spring Harbor Symposia on Quantitative Biology* 51:159-167, 1986.
22. DeBonville, D.A., and Riedle, G.E., "A Robotic Workstation for the Isolation of Recombinant DNA," in *Advances in Laboratory Automation—Robotics*, vol. 3, p. 353.
23. Donahue, R.P., Bias, W.B., Renwick, J.H., et al., "Probable Assignment of the Duffy Blood Group Locus to Chromosome 1 in Man," *Proceedings of the National Academy of Sciences USA* 61:949-955, 1968.
24. Donis-Keller, H., Green, P., Helms, C., et al., "A Genetic Linkage Map of the Human Genome," *Cell* 51:319-337, 1987.
25. Evans, G.A., and Wahl, G.M., "Cosmid Vectors for Genomic Walking and Restriction Mapping," in *Methods in Enzymology: A Guide to Molecular Cloning*, vol. 152, (in press).
26. Foley, B., Nelson, D., Smith, M.T., et al., "Cross-Sections of the Genbank Database," *Trends in Genetics* 2:233-236, 1986.
27. Gall, J.G., letter to the editor, *Science* 233:1367-1368, 1986.
28. Gerhard, D.S., Kawasaki, E.S., Bancroft, F.C., et al., "Localization of a Unique Gene by Direct Hybridization," *Proceedings of the National Academy of Sciences USA* 78:3755-3759, 1981.
29. Gray, J.W., Dean, P.N., Fuscoe, J.C., et al., "High-Speed Chromosome Sorting," *Science* 238:323-329, 1987.
30. Gray, J.W., Langlois, R.G., Carrano, A.V., et al., "High Resolution Chromosome Analysis: One and Two Parameter Flow Cytometry," *Chromosoma* 73:9-27, 1979.
31. Gusella, J.F., Wexler, N.S., Conneally, P.M., et al., "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease," *Nature* 306:234-238, 1983.
32. Hayashibara International Workshop on Automatic and High Speed DNA-Base Sequencing, Hayashibara Biochemical Laboratory, Okayama, Japan, July 7-9, 1987.
33. Hendrix, R.W., Roberts, J.W., Stahl, F.W., et al., *Lambda II* (Cold Spring Harbor, NY: Cold Spring Harbor Press, 1982).
34. Hohn, B., and Collins, J., "A Small Cosmid for Efficient Cloning of Large DNA Fragments," *Gene* 11:291-298, 1980.
35. "Human Gene Mapping 8," *Cytogenetics and Cell Genetics* 40:1-4, 1985.
36. Ish-Horowicz, D., and Burke, J.F., "Rapid and Efficient Cosmid Cloning," *Nucleic Acids Research* 9:2989-2998, 1981.
37. Jeffries, A.J., "DNA Sequence Variants in the γ-δ-β-Globin Genes of Man," *Cell* 1:1-10, 1979.
38. Jones, S.S., Brown, J.E., Vanstone, D.A., et al., "Automating the Purification and Isolation of Synthetic DNA," *BioTechnology* 5:67-70, 1987.
39. Jürgens, G.E., Wieschaus, C., Nüsslein-Volhard, C., et al., "Mutations Affecting the Pattern of the Larval Cuticle in *Drosophila melanogaster* II: Zygotic Loci on the Third Chromosomes," *Roux's Archive of Developmental Biology* 193d:283-295.
40. Kan, Y.W., and Dozy, A.M., "Polymorphism of DNA Sequence Adjacent to Human Beta-Globin Structural Gene: Relationship of Sickle Mutation," *Proceedings of the National Academy of Sciences USA* 75:5631-5635, 1978.
41. Knobeloch, D.W., Hildebrand, C.E., Moyzis, R.K., et al., "Robotics in the Human Genome Project," *Bio/Technology* 5:1284-1287, 1987.
42. Kohara, Y., Akiyama, K., and Isono, K., "The Physical Map of the Whole *E. Coli* Chromosome: Application of a New Strategy for Rapid Analysis and

Sorting of a Large Genomic Library," *Cell* 50:495-508, 1987.

43. Lander, E.S., and Botstein, D., "Mapping Complex Genetic Traits in Humans: New Strategies Using a Complete RFLP Linkage Map," *Cold Spring Harbor Symposia on Quantitative Biology* 51:49-62, 1986.
44. Lander, E.S., and Botstein, D., "Strategies for Studying Heterogeneous Genetic Traits in Humans by Using a Linkage Map of Restriction Fragment Length Polymorphisms," *Proceedings of the National Academy of Sciences USA* 83:7353-7357, 1986.
45. Lange, K., and Boehnke, M., "How Many Polymorphic Genes Will It Take To Span the Human Genome?" *American Journal of Human Genetics* 34:842-845, 1982.
46. Lebo, R.V., Anderson, L.A., Lau, Y.-F.C., et al., "Flow-Sorting Analysis of Normal and Abnormal Human Genomes," *Cold Spring Harbor Symposia on Quantitative Biology* 51:169-176, 1986.
47. Martin, W.J., and Davies, W.R., "Automated DNA Sequencing: Progress and Prospects," *BioTechnology* 4:890-895, 1986.
48. Maxam, A.M., and Gilbert, W., "A New Method for Sequencing DNA," *Proceedings of the National Academy of Sciences USA* 74:560-564, 1977.
49. Maxam, A.M., and Gilbert, W., "Sequencing End-Labeled DNA with Base-Specific Chemical Cleavage," *Methods in Enzymology* 65:499-560, 1980.
50. McKusick, V.A., "The Morbid Anatomy of the Human Genome: A Review of Gene Mapping in Clinical Medicine," *Medicine* 65:1-33, 1986.
51. McKusick, V.A., and Ruddle, F.H., "Toward a Complete Map of the Human Genome," *Genomics* 1:103-106, 1987.
52. National Research Council, *Mapping and Sequencing the Human Genome*, (Washington, DC: National Academy Press, 1988.)
53. Nüsslein-Volhard, C., Wieschaus, E., and Kluding, H., "Mutations Affecting the Pattern of Larval Cuticle in *Drosophila melanogaster* I: Zygotic Loci on the Second Chromosome," *Roux's Archives of Developmental Biology* 193:267-282, 1984.
54. Ohno, S., "An Argument for the Genetic Simplicity of Man and Other Mammals," *Journal of Human Evolution* 1:651-662, 1972.
55. Olson, M.V., Dutchik, J.E., and Graham, M.Y., "Random-Clone Strategy for Genomic Restriction Mapping in Yeast," *Proceedings of the National Academy of Sciences USA* 83:7826-7830, 1986.
56. Pardue, M.L., and Gall, J.G., "Chromosomal Location of Mouse Satellite DNA," *Science* 168:1356-1358, 1970.
57. Rotman, D., "Sequencing the Entire Human Genome," *Industrial Chemist* (December):18-21, 1987.
58. Ruddle, F., Bentley, K.L., and Ferguson-Smith, A., "Physical Mapping Review," contract report to the Office of Technology Assessment, 1987.
59. Saiki, R.K., Sharf, S., Faloona, F., et al., *Science* 230: 1350-1354, 1985.
60. Sanger, F., Nilken, S., and Coulson, A.R., "DNA Sequencing With Chain-Terminating Inhibitors," *Proceedings of the National Academy of Sciences USA* 74:5463-5468, 1980.
61. Schwartz, D.C., and Cantor, C.R., "Separation of Yeast Chromosome-Sized DNAs by Pulsed Field Gel Electrophoresis," *Cell* 37:67-75, 1984.
62. Shera, B., Lawrence Livermore National Laboratory, Livermore, CA, personal communication, February 1988.
63. Smith, C.L., Econome, J.G., Schutt, S., et al., "A Physical Map of the *Escherichia coli* Genome," *Science* 236:1448-1453, 1987.
64. Solomon, E., and Bodmer, W.F., "Evolution of Sickle Variant Gene," *The Lancet* April 28, 1979, p.923.
65. Southern, E.M., "Detection of Specific Sequences Among DNA Fragments Separated by Gel Electrophoresis," *Journal of Molecular Biology* 98:503-517, 1975.
66. Staden, R., "A New Method for Storage and Manipulation of DNA Gel Reading Data," *Nucleic Acids Research* 8:3673-3694, 1980.
67. Stinson, S., "System Automates DNA Amplification," *Chemical and Engineering News*, Dec. 21, 1987, p. 24.
68. U.S. Congress, Office of Technology Assessment, *New Developments in Biotechnology, 4: U.S. Investment in Biotechnology* (Washington, DC: U.S. Government Printing Office, in press).
69. Vogel, F., and Motulsky, A.G., *Human Genetics: Problems and Approaches* (New York: Springer-Verlag, 1986), pp. 369-370.
70. Wada, A., "Automated High-Speed DNA Sequencing," *Nature* 325:771-772, 1987.
71. Waterson, R., Medical Research Council, Cambridge, England, personal communication, October 1987.
72. Watson, J.D., Hopkins, N.H., and Roberts, J.W., *Molecular Biology of the Gene* (Menlo Park: The Benjamin/Cummings Publishing Co., 1987).
73. Watson, J.D., and Crick, F.H.C., "Genetic Implications of the Structure of Deoxyribonucleic Acid," *Nature* 171:964-967, 1953.
74. Watson, J.D., and Crick, F.H.C., "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature* 171:737-738, 1953.
75. Weiss, M., and Green, H., "Human-Mouse Hybrid

Cell Lines Containing Partial Complements of Human Chromosomes and Functioning Human Genes," *Proceedings of the National Academy of Sciences USA* 58:1104-1111, 1967.

76. White, R., and Lalouel, J.-M., "Chromosome Mapping With DNA Markers," *Scientific American* 258:40-48, 1988.
77. White, R., Lippert, M., Bishop, D.T., et al., "Construction of Linkage Maps with DNA Markers for Human Chromosomes," *Nature* 313:101-105, 1985.
78. Wieschaus, E.C., Nüsslein-Volhard, C., and Jürgens, G., "Mutations Affecting the Pattern of the Larval Cuticle in *Drosophila Melanogaster* III: Zygotic Loci on the X-Chromosome and Fourth Chromosome," *Roux's Archive of Developmental Biology* 193:296-307, 1984.
79. Williams, B.G., and Blattner, F.R., "Bacteriophage Lambda Vectors for DNA Cloning," in *Genetic Engineering,* vol. 2, J.K. Setlow and A. Hollaender (eds.), (New York, NY: Plenum Press, 1980).
80. Wilson, E.B., "The Sex Chromosomes," *Arch. Mikrosk. Anat. Entwicklungsmech.* 77:249, 1911.
81. Yunis, J.J., Sawyer, J.R., and Dunham, K., "The Striking Resemblance of High-Resolution G-Banded Chromosomes of Man and Chimpanzee," *Science* 208:1125-1148, 1980.

Chapter 3

Applications to Research in Biology and Medicine

CONTENTS

Page
Introduction 55
Applications in Medicine 56
  Developing Diagnostic Tools 56
  Isolating Genes Associated With Disease 59
  Developing Human Therapeutics 62
  Prospects for Human Gene Therapy 64
Applications in Human Physiology and Development 65
  Identification of Protein-Coding Sequences 65
  Approaches to Understanding Gene Function 66
Applications in Molecular Evolution 68
Applications in Population Biology 72
Chapter 3 References 73

Boxes

Box *Page*
3-A. Why Sequence Entire Genomes? 57
3-B. Duchenne and Becker's Muscular Dystrophies 63
3-C. From Gene Structure to Protein Structure: The Protein-Folding Problem . . . 66
3-D. Constructing the Evolutionary Tree: Morphology v. Molecular Genetics in the Search for Human Origins 70
3-E. The Origin of Human Beings: Clues From the Mitochondrial Genome 71
3-F. Molecular Anthropology 72
3-G. Implications of Genome Mapping for Agriculture 73

Figures

Figure *Page*
3-1. Mapping at Different Levels of Resolution 56
3-2. The Use of Synthetic DNA Probes To Clone Genes 60

Tables

Table *Page*
3-1. Examples of Single-Gene Diseases 57
3-2. Some Companies Developing DNA Probes for Diagnosis of Genetic Diseases 58
3-3. The Size of Human Genes 61
3-4. Some Human Gene Products With Potential as Therapeutic Agents 64
3-5. Classification of Human Proteins by Invention Period 69

Chapter 3

Applications to Research in Biology and Medicine

> "[Physical and genetic maps] will certainly be very useful [but] you have to interpret that sequence, and that's going to be a lot of work. It will be like having a whole history of the world written in a language you can't read."
>
> Joseph Gall
> *American Scientist* 76:17-18,
> February 1988

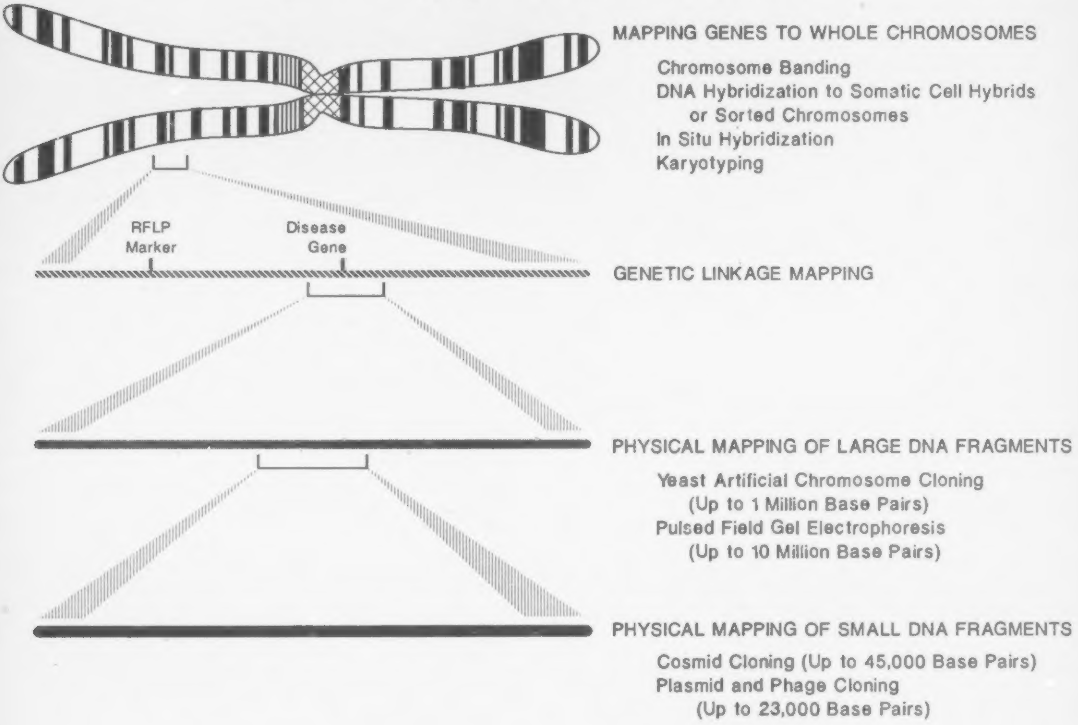
INTRODUCTION

Research efforts aimed at creating genetic linkage and physical maps of chromosomes or entire genomes are collectively referred to in this report as genome projects (see chs. 1 and 2). The goals of genome projects are to develop technologies and tools for mapping and sequencing DNA and to complete maps of human and other genomes. Proponents expect that the products and processes generated from genome projects will enable researchers to answer important questions in biology and medicine. Meeting this objective, however, will depend on the success of concurrent projects aimed at *analyzing* the information generated from mapping genomes. Interpreting genome maps will require the combined efforts of individuals with expertise in structural biology, cell biology, population biology, biochemistry, genetics, computer science, and other fields.

Biology and medicine have already benefited from efforts to map and sequence specific genes from human and other organisms. Some questions might be addressed sooner or better, however, if more extensive genetic linkage maps, cDNA maps, contig maps, and DNA sequences were available (figure 3-1). (See ch. 2 for detailed discussion of the types of genetic linkage and physical maps.) Research on inherited and nongenetic diseases, the physiology and development of organisms, the molecular basis of evolution, and other fundamental problems in biology could all be facilitated in the long run by genome projects.

Scientists continue to debate about which applications depend on information from maps of entire genomes and which require only maps of specific regions. The value of a complete DNA sequence of a reference human genome is the most hotly contended scientific issue (see box 3-A). Focused research has been the mode of molecular genetics to date: Scientists have targeted specific regions of genomes for intensive study. Many of the potential applications of genome mapping summarized in this chapter have already been and will continue to be achieved by targeted research projects. Wherever possible, therefore, this chapter attempts to differentiate between the uses for which extensive maps will be necessary and those for which partial maps are adequate.

Figure 3-1.—Mapping at Different Levels of Resolution


MAPPING GENES TO WHOLE CHROMOSOMES
Chromosome Banding
DNA Hybridization to Somatic Cell Hybrids or Sorted Chromosomes
In Situ Hybridization
Karyotyping
RFLP Marker
Disease Gene
GENETIC LINKAGE MAPPING
PHYSICAL MAPPING OF LARGE DNA FRAGMENTS
Yeast Artificial Chromosome Cloning (Up to 1 Million Base Pairs)
Pulsed Field Gel Electrophoresis (Up to 10 Million Base Pairs)
PHYSICAL MAPPING OF SMALL DNA FRAGMENTS
Cosmid Cloning (Up to 45,000 Base Pairs)
Plasmid and Phage Cloning (Up to 23,000 Base Pairs)

APPLICATIONS IN MEDICINE

Genome projects have accelerated the production of new technologies, research tools, and basic knowledge. At current or perhaps increased levels of effort, they may eventually make possible control of many human diseases—first through more effective methods of detecting disease, then, in some cases, through development of effective therapies based on improved understanding of disease mechanisms. Advances in human genetics and molecular biology have already provided insight into the origins of such diseases as hemophilia, sickle cell disease, and hypercholesterolemia.

The new technologies for genetics research will also help in the assessment of public health needs. Techniques for sequencing DNA rapidly, for example, should permit the detection of mutations following exposure to radiation or environmental agents. Susceptibilities to environmental and work place toxins might be identified as more detailed genetic linkage maps are developed, and special methods of surveillance can be used to monitor individuals at risk. By providing tools for determining the presence or absence of pathogens (e.g., bacteria and viruses) in large numbers of individuals as well as identifying genetic factors that render some human beings more susceptible to infection than others, genome projects might also yield methods for tracking epidemics through populations.

Developing Diagnostic Tools

The use of DNA hybridization probes for detecting changes, such as restriction fragment length polymorphisms (RFLPs), in the DNA of in-

Box 3-A.—Why Sequence Entire Genomes?

Rapid advances in technology have made it feasible to sequence the entire genome of an organism, at least a small one such as bacteria or yeast. Researchers do not yet agree, however, on the value of a complete DNA sequence of a genome the size of the human genome. Several types of arguments have been made in favor of sequencing entire genomes:

- The information in a genome is the fundamental description of a living system—it is what the cell uses to construct a copy of itself—and so is of fundamental concern to biologists.
- Genome sequences provide a conceptual framework within which much future research in biology will be structured. Questions concerning control of gene expression (signals for control of gene expression, genome replication, development mechanisms, and so on) ultimately depend on knowing genome sequences.
- The genomes of some higher organisms, including those of human beings, have repeated DNA sequences, sequences of unknown function, and some sequences which are likely to have no function, comprising nearly 90 percent of the total DNA content. Without the complete DNA sequence of several genomes, it will be impossible to determine whether such sequences have meaning or are ancestral "junk" sequences.
- Genome sequences are important for addressing questions concerning evolutionary biology. The reconstruction of the history of life on this planet, the definition of gene families (also critical to other areas of biology), and the search for a universal ancestor all require an understanding of the organization of genomes.
- Genomes are natural information storage and processing sytems; unraveling them may be of general interest to computer and physical scientists.

Other scientists would argue that these possible applications can be derived from sequences of single genes or larger regions of chromosomes. They believe it is a waste of time and money to sequence the entire human genome, particularly because some regions have no known or essential function. Many of these researchers favor sequencing only those regions believed to be clinically or scientificially important, including expressed sequences and sequences involved in the control of gene expression, and putting the others off indefinitely.

SOURCES:
National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
C. Woese, University of Illinois, Urbana, personal communication, June 1987.

Table 3-1.—Examples of Single-Gene Diseases

Disease	Description	Genetic marker identified	Gene cloned	Protein identified
Duchenne muscular dystrophy	Progressive muscle deterioration	Yes	Yes	Yes
Cystic fibrosis	Lung and gastrointestinal degeneration	Yes	No	No
Huntington's disease	Late-onset disorder with progressive physical and mental deterioration	Yes	No	No
Sickle cell anemia	Deformed red blood cells block blood flow	Yes	Yes	Yes
Hemophilia	Defect in clotting factor VIII causes uncontrolled bleeding	Yes	Yes	Yes
Beta-Thalassemia	Failure to produce sufficient hemoglobin	Yes	Yes	Yes
Chronic granulomatous disease	Frequent bacterial and fungal infections involving lungs, liver, and other organs	Yes	Yes	Yes—tentative
Phenylketonuria	Enzyme deficiency that causes brain damage and mental retardation	Yes	Yes	Yes
Polycystic kidney disease	Pain, hypertension, kidney failure in half of victims	Yes	No	No
Retinoblastoma	Cancer of the eye	Yes	Yes	Yes

SOURCE: Office of Technology Assessment, 1988.

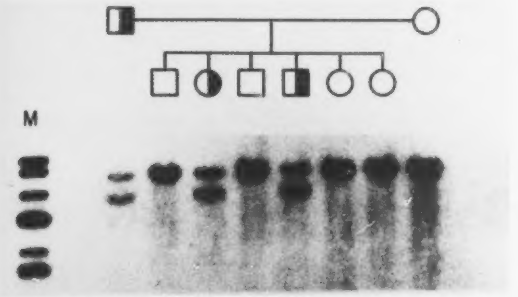


Photo credit: Ray White, The University of Utah Medical Center, Salt Lake City, UT. Reprinted with permission from American Journal of Human Genetics *39:300-306, 1986.*

Identification of a genetic marker showing linkage between high levels of low-density lipoprotein (LDL) cholesterol and the genetic locus for the LDL receptor gene using restriction fragment length polymorphism (RFLP) analysis of the LDL receptor genes from a multigenerational family with inherited hypercholesterolemia. A radioactively labeled DNA fragment from the cloned LDL receptor gene was used as a probe to observe differences among affected and unaffected individuals in the numbers of electrophoretically separated DNA fragments after cutting the DNA with a restriction enzyme. Individuals without the polymorphism are represented as unfilled squares (males) or circles (females) and show only one DNA fragment. Half-filled symbols represent individuals with one allele for the defective gene and one for the normal gene and show two DNA fragments. The lane marked "M" is a set of DNA fragment size markers.

dividuals with genetic diseases is described in detail in chapter 2. Such methods of DNA analysis offer several advantages over traditional approaches to the study of human disease. Knowing the organization of genes on chromosomes and their DNA sequences could enable clinicians to detect mutant genes before a disease manifests itself in the form of damaged cells or tissues and will eventually lead to a more complete understanding of the pathogenesis of human disease [Friedmann, see app. A].

The study of randomly selected RFLP markers in human families has revealed linkages to a number of genetic diseases (table 3-1) (1,3,6, 10,16,17, 23,25,29,32,37,38,41,42,46,52,55,56). As the chromosomal locations of more disease-causing genes are identified, more probes for diagnosing genetic diseases will become available. **A genetic linkage map saturated with RFLP markers (or one with other polymorphic markers) is viewed by many molecular geneticists as crucial to the development of diagnostic reagents for the remaining human genetic diseases** [Friedmann, see app. A]. (See table 3-2 for a list of companies developing diagnostic probes for such diseases.)

It is important to recognize that DNA probes for RFLP markers are not always reliable tools for diagnosing genetic diseases before the onset of symptoms. Without enough data from relatives of potential disease carriers, it may not be possible to confirm the linkage between a particular RFLP marker and a genetic disease. The main limitation to reliable diagnosis of most genetic diseases is the lack of an adequate number of DNA samples from several generations of affected and unaffected individuals.

Many available RFLP markers can be used only in a few families, and the RFLP marker map is a cumulative one that aggregates the data from many families. The largest standard data set is derived from the Center for the Study of Human Polymorphism (CEPH) in Paris (see ch. 7 on international efforts in genome mapping). The data collected by CEPH are taken from 40 families around the world, most of which do not have any known genetic disease. Materials from these families are used to locate RFLP and other polymorphic markers. Once markers have been identified, they can be tested for linkage to a particular genetic disease in families known to have that disease. The

Table 3-2.—Some Companies Developing DNA Probes for Diagnosis of Genetic Diseases

Company	Probes under development
California Biotechnology (Mountain View, CA)	Susceptibility to heart disease
Cetus Corporation (Emeryville, CA)	sickle cell anemia
Collaborative Research (Bedford, MA)	Cystic fibrosis Duchenne muscular dystrophy Polycystic kidney disease
Integrated Genetics (Framingham, MA)	Cystic fibrosis Hemophilia B Huntington's disease Polycystic kidney disease Sickle cell anemia
Lifecodes (Elmsford, NY)	Cystic fibrosis Down's syndrome Polycystic kidney disease Sickle cell anemia

SOURCE: Office of Technology Assessment, 1988.

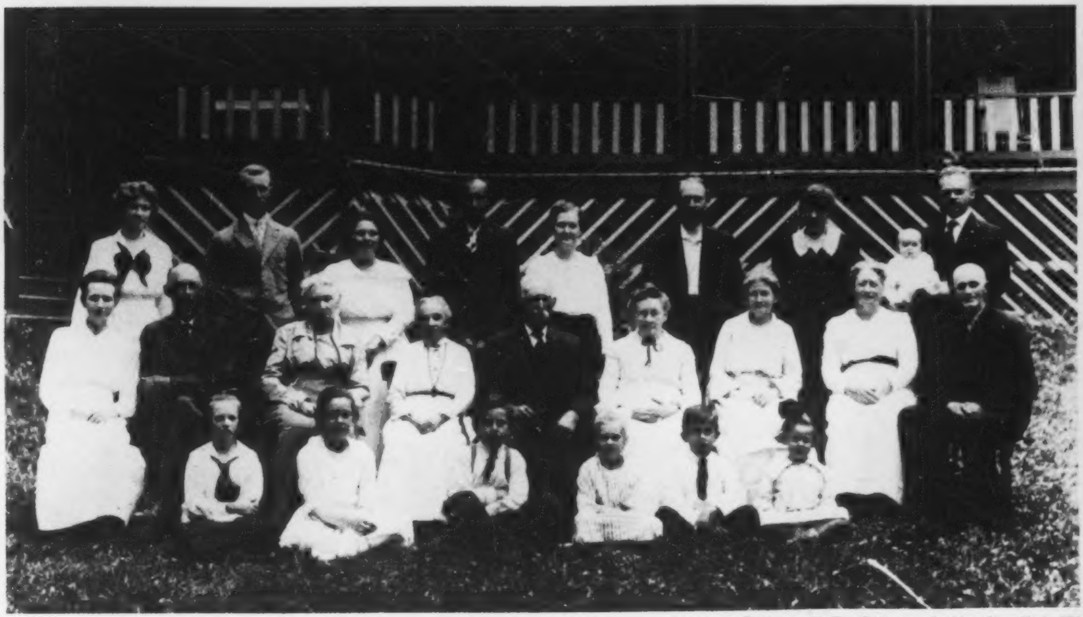


Photo credit: The Bettmann Archive, New York, NY

A large New England family of the early 1900s spanning three generations. Samples of genomic DNA from members of such families are very useful for constructing genetic linkage maps, such as a RFLP map.

CEPH families are large, selected to enable scientists to trace DNA markers through at least three generations.

Isolating Genes Associated With Disease

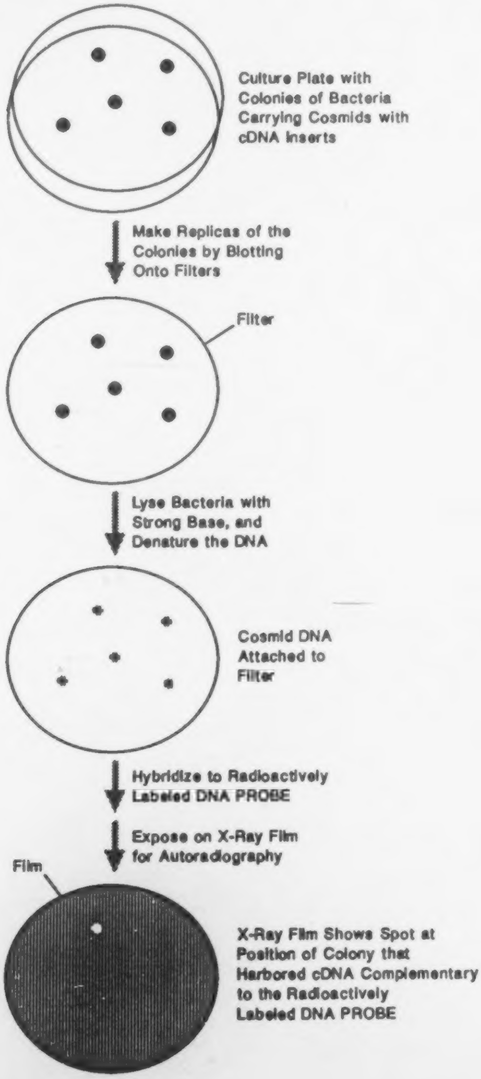
Some inherited human diseases arise from or cause differences in detectable proteins that circulate in the blood, such as human growth hormone and insulin. A research scheme called forward genetics has been used to isolate the genes encoding these proteins. In this strategy, a gene is cloned after the altered protein product has been characterized. Other genetic diseases, such as retinoblastoma, chronic granulomatous disease, and Duchenne muscular dystrophy, involve protein products that were not identified before the corresponding gene was cloned. An experimental approach called reverse genetics was used to find these genes. First the gene containing the mutation responsible for the disease is linked to a RFLP or other polymorphic marker, then the gene and its protein product are isolated and characterized [Friedmann, see app. A].

Forward Genetics

Until recently, most methods for cloning disease-associated genes required prior characterization of the biochemical defect responsible for the disease. Using the forward genetics approach, researchers identify the mutant gene product--a protein—then isolate a clone of the gene from a library of cDNA clones (clones made from DNA copies of the mRNA transcripts of genes—see ch. 2). If the protein has been purified, antibodies can be made and used to select for clones of cells expressing the product. Alternatively, if part of the protein's amino acid sequence is known, synthetic DNA probes complementary to the *exons*, or protein-coding sequences, of the gene can be designed, based on the genetic code (figure 3-2). Once the cDNA clones are isolated, they can be used as DNA probes to pick out clones from genomic libraries.

Figure 3-2.—The Use of Synthetic DNA Probes To Clone Genes

Amino Acid Sequence of a Small Region of a Protein	Methionine	Tyrosine	Arginine	Methionine	Glutamine	Leucine	Serine	Cysteine
An mRNA Sequence Predicted from Amino Acid Codon Usage Frequencies	AUG	UAC	AGG	AUG	CAA	CUG	UCU	UGC
DNA PROBE Sequence Complementary to Predicted mRNA Sequence	TAC	ATG	TCC	ATC	GTT	GAC	AGA	ACG


Culture Plate with Colonies of Bacteria Carrying Cosmids with cDNA Inserts
Make Replicas of the Colonies by Blotting Onto Filters
Filter
Lyse Bacteria with Strong Base, and Denature the DNA
Cosmid DNA Attached to Filter
Hybridize to Radioactively Labeled DNA PROBE
Expose on X-Ray Film for Autoradiography
Film
X-Ray Film Shows Spot at Position of Colony that Harbored cDNA Complementary to the Radioactively Labeled DNA PROBE

The difference between cDNA copies of genes and genes on chromosomes is that the latter have both exons and *introns* (noncoding sequences interrupting protein-coding sequences). The genes in the human genome range from fewer than 1,000 base pairs to more than 2 million base pairs in size and are thus typically too large to be contained on standard cloning vectors (table 3-3). The cDNA clones, which are smaller because they contain only exons, are useful because they can be introduced into bacteria, yeast, or mammalian tissue culture cells and transcribed and translated into protein. The resulting proteins can be used in studies of the physiology of diseases or in some cases as human therapeutics.

The utility of the various types of physical maps in the forward genetics strategy depends on the purpose of isolating the gene. If only cDNA copies of a particular gene are needed for making large quantities of the protein product, then extensive genomic maps would not be necessary. If the cDNA copy of the gene is to be used as a DNA probe for isolating the whole gene from a collection of genomic DNA clones, or for studying the organization of the genome in the region of interest, then a contig map illustrating the order of DNA segments from the relevant portion of the genome would be very useful.

Reverse Genetics

Reverse genetics has made it possible to isolate genes associated with inherited diseases for which no specific biochemical defect has been established. To do this, the genetic disease is usually linked first to a particular chromosome by studying inheritance patterns at the DNA level. The general region of the gene on the chromosome is identified using DNA probes for RFLP markers on that chromosome. Samples of DNA from families of individuals afflicted with the genetic disease are tested with a set of DNA probes which hybridize to markers spaced throughout the chromosome until a linkage between the mutant gene that causes the disease and the RFLP is detected. The location of the gene is then identified more pre-

Table 3-3.—The Size of Human Genes

Gene	Gene size (in thousands of nucleotides)	mRNA size (in thousands of nucleotides)	Number of introns
Small:			
Alpha-globin	0.8	0.5	2
Beta-globin	1.5	0.6	2
Insulin	1.7	0.4	2
Apoliproprotein E	3.6	1.2	3
Parathyroid	4.2	1.0	2
Protein kinase C	11.0	1.4	7
Medium:			
Collagen I			
Pro-alpha-1(I)	18.0	5.0	50
Pro-alpha-2(I)	38.0	5.0	50
Albumin	25.0	2.1	14
High-mobility group CoA reductase	25.0	4.2	19
Adenosine deaminase	32.0	1.5	11
Factor IX	34.0	2.8	7
Catalase	34.0	1.6	12
Low-density-lipoprotein receptor	45.0	5.5	17
Large:			
Phenylalanine hydroxylase	90.0	2.4	12
Factor VII	186.0	9.0	25
Thyroglobulin	300.0	8.7	36
Very large:			
Duchenne muscular dystrophy	2,000.0	17.0	50

SOURCE: Victor McKusick, The Johns Hopkins University, Baltimore, MD.

cisely by using additional probes for closely spaced markers that cover the region of interest.

In order to distinguish the gene locus that actually causes the disease from nearby, but unrelated, genes, it is generally necessary to demonstrate that the identified gene is expressed abnormally in tissues from patients with the disease. A genomic region 1 million base pairs in length, for example, could contain as many as 100 genes. In such cases, it is necessary to use biochemical methods to identify the gene that is responsible for the disease. Techniques for detecting messenger RNA transcripts or proteins can be used to search for differences in amounts of *gene product* in the tissues of affected and unaffected individuals; these differences can then be correlated with an alteration in a particular gene. Retinal cells were analyzed in this way as part of the search for the retinoblastoma gene (18,28), as were muscle cells in individuals with and without Duchenne muscular dystrophy (see box 3-B). Once the gene product has been identified, it is possible to study the physiology of a particular disease with the aim of identifying a therapy or preventive treatment.

Although reverse genetics is generic in concept, the amount of effort involved in isolating and characterizing genes using genetic and physical map data varies. Over 100 person-years have been spent searching for the gene that causes cystic fibrosis—an effort that has led to localizing the gene on a small region of chromosome 7 but not to finding the gene itself or determining how it causes the disease (17). On the other hand, researchers identified and isolated the gene for chronic granulomatous disease in far fewer person-years (38). The existence of DNA probes for RFLP markers has also made possible the identification of the genes for Duchenne muscular dystrophy (32) and retinoblastoma (18,28) [Friedmann, see app. A].

The technical difficulty involved in locating the gene responsible for a particular disease by reverse genetics usually depends on the physical map distance between the nearest RFLP marker and the linked gene. Existing RFLP maps of the human genome have a resolution of only about 10 centimorgans (approximately 10 million base pairs). **A map with markers spaced every 1 centimorgan would make it much less time-consuming to locate the genes by reverse genetics** (7,33). Such a map would be constructed using a pool of several thousand DNA probes that detect RFLP markers spaced about every 1 million base pairs throughout the genome. **A library of clones made from overlapping segments of the genome and a contig map illustrating the relative position of each clone with respect to its neighbors would also be useful in reverse genetics.** These tools would spare researchers the labor-intensive step of isolating and characterizing all of the genomic clones between the marker and the gene of interest; the only work remaining would be to associate the characteristics of the disease with the correct clone or clones.

Identification of Genes Involved in Polygenic Disorders

Genetic linkage maps of the human genome are also useful for characterizing inherited diseases caused by more than one factor, often referred to as *polygenic disorders*. Among the diseases for which more than one gene is likely to be responsible are certain cancers, diabetes, and coronary heart disease (27). For example, in a complex disorder such as coronary heart disease, blood plasma lipoproteins, the coagulation system, and elements of the arterial walls all play a role, so the number of genes involved can be very large (40). Some scientists argue that the RFLP maps currently available, with markers spaced an average of 10 centimorgans apart, are a sufficient starting point for studies of polygenic diseases (11). **Higher-resolution RFLP maps, such as a 1-centimorgan map, would no doubt simplify the job of identifying the genes responsible for polygenic disorders.**

Developing Human Therapeutics

Forward genetics has yielded important results in the area of drug development. As stated earlier, the ability to use cDNA clones has been crucial to the development of commercial products such as human growth hormone and insulin and to potential human therapeutics such as tumor necrosis factor and interleukin-2, therapeutics that would not otherwise be available in the quantity or quality necessary for effective use (table 3-4)

Box 3-B.—Duchenne and Becker's Muscular Dystrophies

Duchenne muscular dystrophy (DMD) is a genetic disease that affects 1 in 3,000 to 1 in 3,500 male infants born. Becker's muscular dystrophy is a similar but milder disorder with much lower incidence. Both diseases begin in childhood and lead to muscle wasting. DMD typically results in death before age 20. The search for the gene causing these diseases and the protein encoded by that gene has been an exciting story of molecular biology in the 1980s. The effort in many ways typifies modern genetics, with extensive international collaboration, study of nonhuman species, and creative use of molecular methods.

The gene causing these diseases had been known for some time to be on the X chromosome because of inheritance patterns. Duchenne and Becker's muscular dystrophies affect primarily boys, who have only one X chromosome, inherited from their mothers. Girls have two X chromosomes and therefore must, as a rule, receive abormal genes from *both* parents in order to develop Duchenne or Becker's muscular dystrophy—a much less likely occurrence.

The search for the gene started with studies of families. DNA from persons with DMD, including several girls and one boy, was collected in an effort to find a common area of the X chromosome that had been lost or altered. Once the correct region of the X chromosome had been identified (its absence was found to cause DMD), DNA from that region was obtained and cloned. The clones were used as DNA probes for complementary mRNA sequences in muscle tissue from affected and unaffected individuals. The purpose was to identify the mRNA gene transcript that was present in unaffected individuals but altered in persons with DMD. The mRNA was located and subsequently shown to encode a large protein called dystrophin found in muscle cells.

The DMD search has uncovered some extraordinary facts. Duchenne and Becker's muscular dystrophies are caused by different changes in the same gene. That gene is the largest found to date, spanning over 2 million base pairs (table 3-3). It is broken into at least 60 exons.

The scientific collaboration that led to the discovery of dystrophin was notably efficient. One paper alone listed 77 authors from 24 research institutions in 8 countries. Molecular probes, clones, and materials from affected patients were openly exchanged, hastening researchers in their quest for the culprit gene.

SOURCES:

K.H. Fischbeck, A.W. Ritter, D.L. Tirschwell, et al., "Recombination With pERT87 (DXS164) in Families With X-Linked Muscular Dystrophy," *Lancet* 2(July):104, 1986.

E.P. Hoffman, A.P. Monaco, C.C. Feenel, et al., "Conservation of the Duchenne Muscular Dystrophy Gene in Mice and Humans," *Science* 238:347-350, 1987.

E.P. Hoffman, R.H. Brown, and L.M. Kunkel, "Dystrophin: The Protein Product of the Duchenne Muscular Dystrophy Locus," *Cell* 51:919-928, 1987.

M. Koenig, E.P. Hoffman, C.J. Bertelson, et al., "Complete Cloning of the Duchenne Muscular Dystrophy (DMD) cDNA and Preliminary Genomic Organization of the DMD Gene in Normal and Affected Individuals," *Cell* 50:509-517, 1987.

L.M. Kunkel et al., "Analysis of Deletions From Patients With Becker and Duchenne Muscular Dystrophy," *Nature* 322:73-77, 1986.

A.P. Monaco, R.L. Neve, C. Colletti-Feener, et al., "Isolation of Candidate cDNAs for Portions of the Duchenne Muscular Dystrophy Gene," *Nature* 323:646-650, 1986.

A.P. Monaco, C.J. Bertelson, C. Colletti-Feener, et al., "Localization and Cloning of Xp21 Deletion Breakpoints Involved in Muscular Dystrophy," *Human Genetics* 75:221-227, 1987.

G.J. van Omenn, J.M. Verkerk, M.H. Hofker, et al., "A Physical Map of 4 Million Base Pairs Around the Duchenne Muscular Dystrophy Gene on the X Chromosome," *Cell* 47:499-504, 1986.

(53). The cDNA clones isolated by forward genetics could be used to make a cDNA map that illustrates the chromosomal locations of expressed regions of DNA. **This cDNA map, plus a library of previously ordered clones of genomic DNA, would be valuable tools for studying the role of certain genes in the manifestation of disease.** Knowledge of the mechanisms directing normal cellular functions will probably lead to important sources of new therapies for human diseases: natural human proteins made from isolated human genes, engineered proteins, and conventionally synthesized drugs designed from a knowledge of the structure of the proteins they target. **Advances in the development of human therapeutic products will be made more rapidly if research in the areas of protein engineering, the relationship of protein structure to function, rational drug design, and others parallels genome mapping efforts.**

Photo credit: Nancy Wexler, Columbia University, New York, NY

A Venezuelan man with Huntington's disease, a rare, late-onset genetic disease that causes degeneration of nerve cells in the brain.

Prospects for Human Gene Therapy

Clinical use of human genetic linkage and physical maps, now largely restricted to diagnosis, may eventually include the insertion of normal DNA directly into human cells to correct a particular genetic defect (54). This practice is called *human gene therapy*. Advances in gene therapy will depend on development of ways to insert DNA into cells safely and to ensure that the inserted DNA corrects the defect (54). Gene mapping will not improve gene therapy directly, and for most diseases the ability to make a diagnosis will precede the availability of an effective treatment. The knowledge gained through use of gene maps will, however, enhance knowledge about the function of genes and thus indirectly improve the prospects for gene therapy [Friedmann, see app. A].

Table 3-4.—Some Human Gene Products With Potential as Therapeutic Agents

Gene product	• Actual or potential therapeutic application
Atrial Natiuretic Factor	• Possible applications in treatment of hypertension and other blood pressure disorders, and for some kidney diseases affecting excretion of salts and water.
Alpha Interferon[a]	• Approved for treatment of hairy cell leukemia; possible broader applications in other cancers.
Beta Interferon	• Inhibits viral infections and may be useful as an anticancer treatment.
Epidermal Growth Factor	• Expected to have applications in wound healing, including burns, and for cataract surgery.
Erythropoietin	• Anticipated treatment use for anemia resulting from chronic kidney disease.
Factor VIII:C[b]	• Prevents bleeding in patients with hemophilia A after injury.
Fibroblast Growth Factor	• Possible use in wound healing and treating burns.
Gamma Interferon	• Possible treatment for scleroderma and arthritis.
Granulocyte Colony Stimulating Factor	• Possible treatment for Acquired Immune Deficiency Syndrome (AIDS) and leukemia.
Human Growth Hormone[a]	• Approved as a treatment for childhood dwarfism; expected to have broader therapeutic potential in treatment for short stature resulting from Turner's syndrome and for wound healing.
Insulin[a]	• Approved for treatment of diabetes.
Interleukin-2	• Possible treatment for various cancers.
Macrophage Colony Stimulating Factor	• Potential applications are for treatment of infectious diseases, primarily parasites; possible cancer therapy.
Superoxide Dismutase	• Possible preventive treatment for damage caused by oxygen-rich blood entry into oxygen-deprived tissues (e.g., during organ transplants).
Tissue Plasminogen Activator	• Approved as treatment for dissolving blood clots associated with heart attacks.
Tumor Necrosis Factor	• Possible anti-tumor therapy.

[a] Approved for commercial sale in the United States by the Food and Drug Administration.

[b] Non-recombinant DNA version has been approved for sale, but the cloned gene product has not.

SOURCE: Office of Technology Assessment, 1988.

APPLICATIONS IN HUMAN PHYSIOLOGY AND DEVELOPMENT

Studies aimed at understanding the molecular basis of inherited diseases may yield information that can be generalized to other physiological processes. Knowledge of the structure and function of genes associated with Alzheimer's disease, for example, might give important clues to the cellular mechanisms regulating aging of brain tissue.

The organization of genes in genomes is another fundamental issue in biology. Is it important for genes to exist on a particular chromosome in a particular order? Comparisons of physical maps of the chromosomes of higher organisms could shed some light on the extent to which gene organization is associated with gene expression and gene function.

The nucleotide sequences of human genes have been and will continue to be important research tools for understanding the basic cellular processes underlying physiology and development. Nevertheless, knowing the DNA sequences of genes and how they translate into the amino acid sequences of protein products is not sufficient to establish how such genes are controlled or how the gene products function in a particular cell or in the organism as a whole. The genetic code that guides the translation of DNA sequence into protein sequence offers only the first step in unraveling the mysteries of the human genome. Understanding the relationship between protein structure and function is the crucial next step, but it faces the greatest number of technical bottlenecks (see box 3-C).

Identification of Protein-Coding Sequences

Individual efforts to clone particular genes will not be eliminated by the availability of genetic linkage and physical maps; rather, they will be redirected toward localizing a particular gene within a region of a chromosome or within the DNA sequence of that region. Because human genes are more often interrupted by introns, the identification of the exon and regulatory sequences in and around genes has proved more difficult in human beings than in lower organisms. The most reliable method of identifying exons is to know the amino acid sequence of the protein product and, using the genetic code, find the corresponding DNA sequence by inspecting the whole gene sequence. DNA sequences can be determined at a faster rate than proteins can be isolated and sequenced, however, so computer-assisted methods offer a more practical approach.

There is a variety of computer software available for predicting exon sequences, some of it more reliable than others (12,14,48) [Mount, see app. A]. As more DNA sequences become available, methods for predicting exons can continue to be refined. **Computer scientists argue that the analysis phase of whole genome sequencing projects will progress efficiently only if the development of new computational and other theory-based predictive methods that can accommodate large sequences is emphasized.**

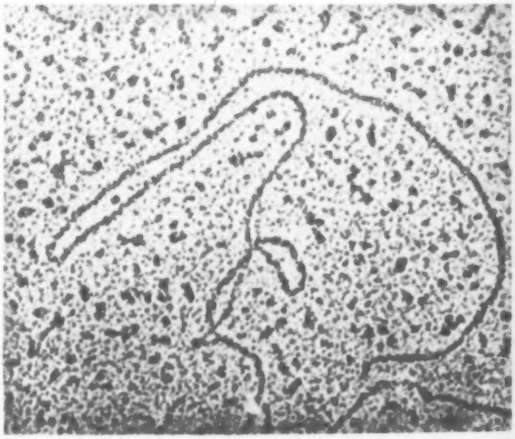


Photo credit: Shirley Tilghman, Princeton University, Princeton, NJ

Electron micrograph revealing an intron sequence interrupting the protein-coding sequences of the mouse beta-globin gene. DNA containing the gene (including intron sequences) was allowed to hybridize (base pair) with beta-globin mRNA that had been isolated from cells in its mature form with no intron sequences. A loop appears in the region of the intron where no complementary sequences exist between the two molecules (see arrow).

Box 3-C.—From Gene Structure to Protein Structure: The Protein-Folding Problem

"Protein-folding is the genetic code expressed in three dimensions," according to Fetrow and co-workers. How does the linear sequence of amino acids code for a protein's structure? How does the three-dimensional conformation of a protein drive its function? Sometimes the amino acid sequence of a protein with an unknown function is similar to that of a protein with a known function; in many such cases, the similarity is a valid indicator of comparable jobs. In other cases, the three-dimensional structure of a protein (the amino acid sequence folded into the actual structure of the protein) gives more reliable clues about function. It is therefore important to develop experimental and theoretical means for determining the three-dimensional structures of proteins. Because proteins are so large, often consisting of multiple domains (discrete portions) with different functions, this generally involves analysis of how each part of a protein contributes to its overall structure.

There is experimental evidence that certain structural domains can serve similar functions in a number of different proteins. It is the combination of domains that gives a protein its unique overall function. A stretch of amino acids in one protein can be nearly identical in sequence to that in another protein, but if the surrounding amino acid sequences are different, then the sequences might fold into domains with quite different three-dimensional structures. At present, scientists cannot predict with certainty how the linear sequence of amino acids in a protein will fold into the protein's three-dimensional structure—thus the protein-folding problem. As genome mapping projects make more gene sequences available, the problem will take on even greater significance. The National Academy of Sciences in a recent report called protein folding "the most fundamental problem at the chemistry-biology interface, and its solution has the highest long-range priority."

Most predictions of three-dimensional structure are based on theories of the behavior of amino acids in certain chemical and physical environments and on information gleaned from viewing the atomic structures of proteins through X-ray diffraction. (X-ray diffraction of protein crystals is an important tool in structural biology—the field dedicated to the study of proteins and other macromolecular structures. It is the most important technique for determining the three-dimensional structure of large proteins at the atomic level.) Existing methods for predicting structure are not reliable for all proteins or protein domains, because structural data are available for only about 200 proteins and for even fewer classes of proteins. There are few membrane proteins in the structure database, for example, and thus little experimental basis for testing predictions about how such important proteins will fold. More structures of proteins need to be determined, using X-ray crystallographic and other biophysical technologies, in order to provide a solid foundation for protein-folding theories. Once the protein-folding problem is solved, the road from gene sequence to gene function will be considerably shortened and will lead, in some cases, toward the development of promising new human therapeutic products.

SOURCES:
T. Blundell, B.L. Sibanda, M.J.E. Sternberg, and J.M. Thornton, "Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules," *Nature* 326:347-352, 1987.
J.S. Fetrow, M.H. Zehfus, and G.D. Rose, "Protein Folding: New Twists," *Bio/Technology* 6:167-171, 1988.
T. Koetzle, Brookhaven National Laboratory, Upton, NY, personal communication, March 1987.
National Academy of Sciences, *Research Briefings* (Washington, DC: 1986, National Academy Press, 1986).

Approaches to Understanding Gene Function

Isolating a gene is not nearly as difficult as determining how the gene and its products function in the cell. The following are some experimental approaches to solving this problem:

- to modify or inactivate the normal function of a gene by replacing it with a modified version,
- to inhibit the function of a gene's mRNA or protein product by using antibodies to the protein or an RNA complementary to the mRNA, and

- to compare the DNA sequence of a cloned gene of unknown function with those of genes whose functions are known.

Using the first two methods, scientists have studied the function of gene products by identifying alterations in the biochemical or physical characteristics of the affected cell or organism (2,15, 24,26,30,36,39,44,50,51). The third strategy is theoretical, using sequence data accumulated from previously characterized gene products to predict a function for a newly identified gene. Such predictions can then be tested experimentally.

Probably the most widely used first step in determining the role of a gene is to find similarities between its DNA sequence and those of genes from other organisms. Yeast, for example, shares with animal cells many of the molecules and processes that are being studied intensively in modern cell biology, including the factors modulating cell structure and dynamics, the components of the machinery that modulates protein secretion from cells, the constituents of basic chemical pathways, and analogs of several mammalian *oncogenes* (genes involved in controlling the rate of cell growth). Many of these factors and processes were first identified or characterized, or both, in higher organisms, but the application of them to yeast genetics has provided new insights (57).

A recent study reported the use of yeast cells to isolate a human gene that can substitute for a yeast gene in regulating the yeast cell's life cycle (34). Plasmid vectors carrying cDNA were introduced into yeast cells to find a human gene product that was similar enough to a yeast gene to replace it in the regulation of the yeast cell's life cycle. This was accomplished by mutating the yeast's copy of the gene and then finding cells that, upon introduction of the appropriate human gene, appeared to regain their normal function. The human cDNA clone identified by this technique is thus a candidate for a protein that regulates life cycles in human cells. Studies of the genomic version of the human gene will be necessary to definitively establish the role of this product in human cell cycle regulation. This example illustrates how yeast genetics and biochemistry can be used to identify human genes with important functions (57).

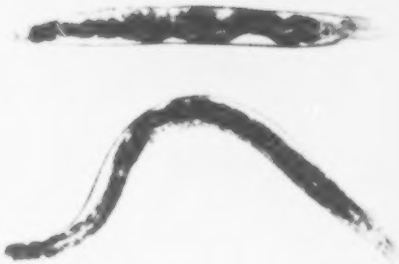


Photo credit: Donald Riddle, University of Missouri, Columbia, MO

Micrograph comparing the appearance of a short, fat nematode mutant called "dumpy" (above) with that of a normal nematode (below). The mutant grows to only two-thirds of the normal body length because of a mutation in a gene for a type of collagen (a protein) needed for normal development (magnified 80 times).

In fruit flies, genetic research and the tools of recombinant DNA have made it clear that certain DNA sequences are involved in regulating the development of the organism. Different sets of genes appear to be expressed at different times in the course of development, causing the patterns observed in the developing embryo. How gene expression is regulated to create developmental patterns is a central question in biological studies of many organisms. Fruit flies are easy to dissect and to manipulate genetically, and much is known about their development; they have therefore proven to be a very useful model. One DNA sequence, called the *homeo box*, was first identified in the genes of fruit flies and later in those of higher organisms, including human beings (31). There is substantial evidence that the homeo box, a short stretch of nucleotides of nearly identical sequence in the genes that contain it, determines when the expression of particular groups of genes is turned on and off during development of the fruit fly (35). As more gene sequences from fruit flies, human beings, and other organisms are determined, more knowledge about the signals governing developmentally expressed genes is likely to be acquired [Mount, see app. A].

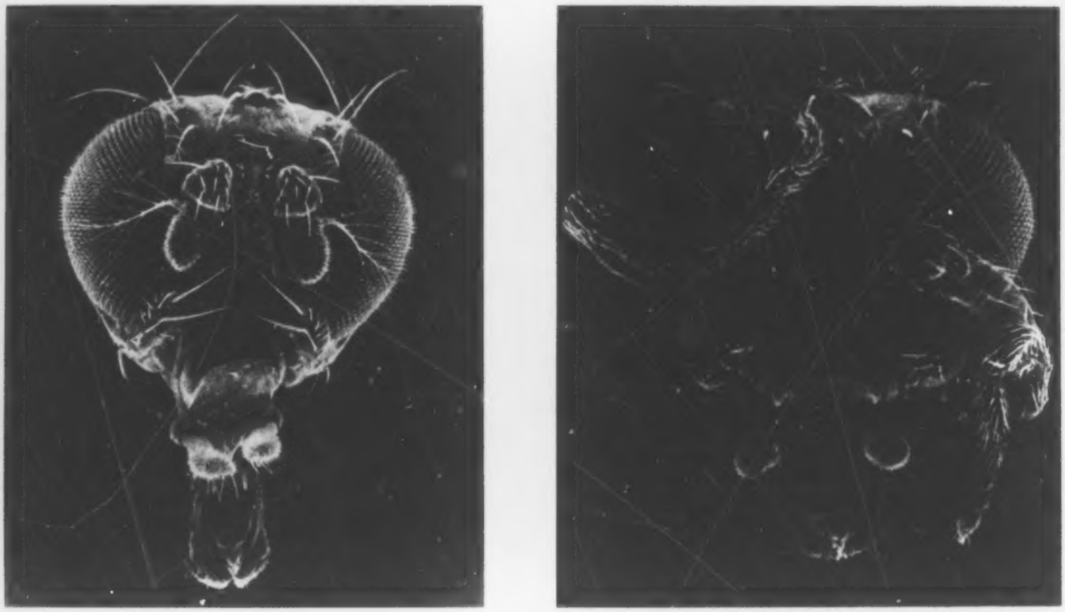


Photo credit: John Postlethwait, University of Oregon, Eugene

Mutations in the gene *Antennapedia*, a homeotic gene, cause the fruit fly to develop an extra pair of antennas. Pictured at left is the normal fruit fly, and at right a fly with the mutation. Homeotic genes have counterparts in humans and vertebrates; each gene has a characteristic DNA sequence within its protein-coding sequences called the homeo box.

APPLICATIONS IN MOLECULAR EVOLUTION

The disciplines of population biology, genetics, molecular biology, and cellular biology merge in the study of how species evolve, constituting the field of molecular evolution. The construction of a physical map of the human genome will permit molecular analysis of several questions fundamental to evolution, including how genomes change and what factors cause them to change, as well as how small-scale changes relate to the overall evolution of the organism (45).

Species with different degrees of relatedness can be usefully compared because their genes, and thus the proteins encoded by those genes, will have differing rates of sequence divergence. The course of human evolution can be read in the sequences of proteins (14). Comparisons of human and mouse DNA sequences are probably the most useful in the identification of genes unique to higher organisms because mice genes are more homologous to human genes than are the genes of any other well-characterized organism. Comparisons of human DNA sequences with those of lower organisms such as the fruit fly or nematode are most useful in the identification of genes encoding proteins that are essential to all multicellular organisms. Finally, since yeasts are single-celled *eukaryotes* (cells whose chromosomes are contained in nuclei), their sequences are most useful in the identification of genes that make proteins whose functions are essential to the life of all eukaryotic cells because such proteins would be least likely to have undergone major changes in the course of their evolution [Mount, see app. A]. Table 3-5 shows how human proteins can be classified by their period of invention: from ancient, to middle-age, to modern (14).

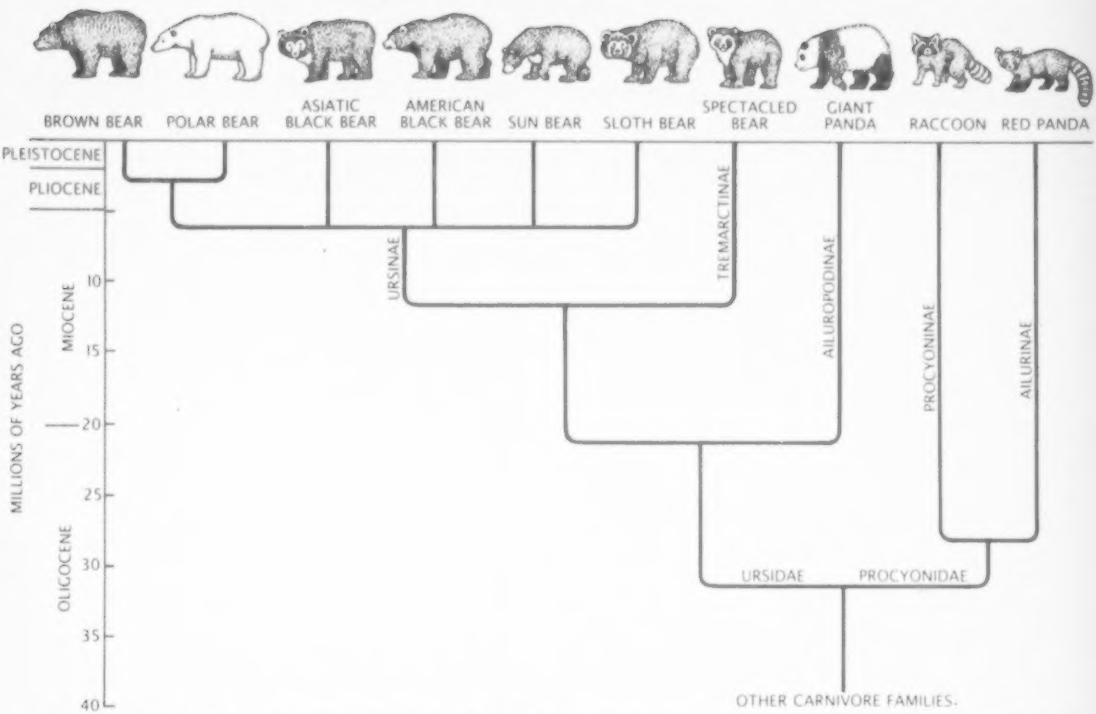

BROWN BEAR
POLAR BEAR
ASIATIC BLACK BEAR
AMERICAN BLACK BEAR
SUN BEAR
SLOTH BEAR
SPECTACLED BEAR
GIANT PANDA
RACCOON
RED PANDA
PLEISTOCENE
PLIOCENE
MIOCENE
OLIGOCENE
MILLIONS OF YEARS AGO
10
15
20
25
30
35
40
URSINAE
TREMARCTINAE
AILUROPODINAE
PROCYONINAE
AILURINAE
URSIDAE
PROCYONIDAE
OTHER CARNIVORE FAMILIES

Photo credit: Stephen O'Brien, The National Cancer Institute, Frederick, MD. Reprinted with permission from Scientific American, November 1987, pp. 102-107.

A phylogenic tree based on data obtained from modern molecular genetic methods places the giant panda in the Ursidae, or bear family. The red panda is left in the Procyonidae, or raccoon family. Molecular analysis of the chromosomes of these pandas suggests that the raccoon and bear families diverged from a common carnivorous ancestor about 35 to 40 million years ago.

Table 3-5.—Classification of Human Proteins by Invention Period

- I. Ancient proteins
  - A. *First editions.* Direct-line descendancy to human and contemporary prokaryotes. Mostly enzymes involved in metabolism.
  - B. *Second editions.* Homologous sequences in human and prokaryotic proteins, but apparently different functions.
- II. Middle-age proteins
  Proteins found in most eukaryotes but prokaryotic counterparts are as yet unknown.
- III. Modern proteins
  - C. *Recent vintage.* Proteins found in animals or plants but not both. Not found in prokaryotes.
  - D. *Very recent inventions.* Proteins found in vertebrate animals but not elsewhere.
  - E. *Recent mosiacs.* Modern proteins clearly the result of shuffling exons.

SOURCE: Adapted from Doolittle, R.F., Feng, D.F., Johnson, M.S., and McClure, M.A., "Relationships of Human Protein Sequences to Those of Other Organisms," *Cold Spring Harbor Symposia on Quantitative Biology* 51:447-456, 1986.

Physical map and sequence data accumulated from many species over the past 10 years have led scientists to recognize patterns of genome change quite different from those proposed earlier. Now, molecular evolutionists are beginning to understand such patterns as the duplication and acquisition of new genes and their corresponding functions, differences in the use of the genetic code among different organisms (48), and differences in the occurrence of gene families in different species (45).

Important questions in molecular evolution arise from the fact that the genes of *prokaryotes* (organisms without nuclei, e.g., bacteria), as well as many genes in yeast and some multicellular organisms, are not interrupted by introns:

- Are the introns found in genes today descendants of extra or unused DNA from bacteria and eukaryotes such as yeast?
- Did prokaryotes rid themselves of intron sequences, or did they never have them?
- How did intron sequences get into the genes that code for modern proteins (14)?

By sequencing similar genes from many species, scientists have found that some introns have been in place for very long evolutionary periods and that the positions of introns within genes divide the genes in ways that correspond to the distinct functional domains of the proteins' structures (5,22,49). These observations have led to new models of molecular evolution (8,13,19-21,47). *The availability of more gene sequence data should facilitate the assessment of theories about the evolution of genes and gene structures.*

Sequences of more human genes, high-level understanding of variations in genomic organization among individuals, and analyses of differences between human beings and other organisms should aid in the evaluation of molecular evolutionary theories on how species originate (see boxes 3-D, 3-E, and 3-F). Recognition of differences in rates of nucleotide substitution, recombination, and other mechanisms responsible for variation in the human genome will lead to a better understanding of the molecular basis of these processes and of the constraints on each. Evaluation of proposed models for the propagation and evolution of multigene families, such as certain classes of cell surface receptors, requires a detailed knowledge not only of the relatedness of the DNA sequences in these genes, but also of their locations in the genome and the DNA sequences of the regions surrounding them (45).

Box 3-D.—Constructing the Evolutionary Tree: Morphology v. Molecular Genetics in the Search for Human Origins

Ever since Linnaeus, biologists have classified animals according to similarities and differences in form and structure. When the concept of evolution took root, these morphological features were used to establish phylogenies—trees or lineages that indicate the evolutionary relationships among species. New and sophisticated methods of genetic analysis have challenged morphology as the prime determinant of family trees. Recent debates about human origins have revealed the potential power of genetic techniques for evolutionary studies.

For the past two decades or so, anthropologists and biologists studying the problem of primate evolution have agreed that chimpanzees and gorillas are closely related enough to be classified in the same family, while humans stand alone in a separate, more distant family. Morphological evidence favors this view. Both chimps and gorillas, for example, walk on their knuckles; humans do not, and the fossils of their most direct ancestors show no features associated with knuckle walking. Chimps and gorillas also share similarities in the thickness and structure of their tooth enamel which suggest a common ancestry separate from humans.

Analyses of the DNA of chimps, apes, and human beings contradict this view. Scientists recently examined comparable segments of DNA in the region of the beta-globin gene from human beings, chimpanzees, gorillas, and orangutans. They sequenced 4,900 base pairs of DNA from this region in each organism, then appended data for nearby regions for which sequences had already been published. In all, they compared a 7,100-base-pair region and concluded that chimpanzee and human gene sequences were the least divergent. The most parsimonious explanation of the data was that human and chimpanzee are more closely related to each other than either is to the gorilla.

The beta-globin study, while it strongly suggests that chimpanzees are the closest cousins of human beings, does not conclusively end the search for human origins. Contradictions remain in the evidence gathered from comparative anatomy and from genetic analysis; studies of other gene loci will be necessary to settle the matter.

SOURCES:
R.L. Cann, "In Search of Eve," *The Sciences* (September/October):30-37, 1987.
R. Lewin, "My Close Cousin the Chimpanzee," *Science* 239:273-275, 1987.
M.M. Miyamoto et al., "Phylogenetic Relations of Humans and African Apes From DNA Sequences in the Globin Region," *Science* 239:369-373, 1987.

Box 3-E.—The Origin of Human Beings: Clues From the Mitochondrial Genome

For more than a century, archaeologists, anthropologists, and biologists have been digging through layers of dirt and rock, sieving fossils and artifacts, in an attempt to figure out when, where, and how human beings differentiated from other primates to become a unique species. These scientists have relied on a variety of tools, everything from the picks and axes used to dig up fossils to sophisticated techniques for determining the age of the bones they have unearthed. Unfortunately, archaeological digs do not always yield perfect clues: Even well-preserved fossil remains are generally incomplete, and there are still missing links, cases in which fossils that could hint at the genealogy of several precursor species have not been found. Thus, it has been difficult to determine exactly when human beings diverged from prehistoric ancestors to become the species now known as *Homo sapiens*.

The development of molecular genetic techniques for analyzing DNA offers a new source of evidence in the ongoing debate about human origins. Techniques for mapping and sequencing DNA allow researchers to compare different species and different individuals from the same species at the most basic level. These comparisons can aid evolutionary studies.

One promising approach is the study of the DNA sequences of mitochondria, small structures that are found in the cells of all multicellular organisms. Mitochondria are the power plants of eukaryotic cells. They produce energy for life processes by providing a site for the combination of oxygen and food molecules. Without them, cells would depend on less efficient processes of energy production and could not survive in an environment containing oxygen. Mitochondria have much in common with bacteria: They are similar in size and shape, they both contain DNA, and they each reproduce by dividing in two.

The DNA in mitochondria can be more useful for some evolutionary studies than the DNA in cell nuclei, for several reasons. First, since it lies outside the cell nucleus and sexual recombination occurs only within the nuclei of sperm and egg cells, mitochondrial DNA is not recombined during sexual reproduction. It is inherited only from the mother. Consequently, changes in the nucleotide sequences are due only to mutation and not to the natural shuffling of DNA that occurs during reproduction. Second, DNA in the mitochondria is not protected as well as DNA in the nucleus, nor does it have the same kinds of mechanisms for repair. Thus, mitochondrial DNA mutates about 10 times as fast as the chromosomal DNA in the cell's nucleus, which means that the mitochondrial genome has evolved more rapidly than the chromosomal genome. Finally, mitochondria are relatively small: They contain approximately 16,000 base pairs, considerably fewer than the 3 billion base pairs in the entire set of human chromosomes, making them easier to analyze.

These three characteristics of mitochondrial DNA—absence of sexual recombination, a high natural mutation rate, and small size—have helped scientists construct a "molecular clock" that can be used to help establish the approximate time and place of human origins. By calculating the rate at which mitochondrial DNA changes and then comparing the DNA sequences of mitochondria from many individuals, researchers have begun to formulate genealogical trees. For example, scientists sequenced samples of mitochondrial DNA from 140 people around the world and used the information to propose that the first *Homo sapiens* lived 200,000 years ago on the African continent. Prior to these findings, anthropologists speculated that human beings originated nearly 1 million years ago. Debate continues among scientists about the validity and proper application of mitochondrial DNA sequences in evolutionary studies, but it is clear that molecular genetics will play a growing role in this area.

SOURCES:

B. Alberts, D. Bray, J. Lewis, et al., "The Evolution of the Cell," in *Molecular Biology of the Cell* (New York, NY: Garland Publishing, 1983).
R.L. Cann, "In Search of Eve," *The Sciences* (September/October):30-37, 1987.
R.L. Cann, M. Stoneking, and A.C. Wilson, "Mitochondrial DNA and Human Evolution," *Nature* 325:31-36, 1986.
R. Lewin, "Molecular Clocks Turn a Quarter Century," *Science* 239:561-563, 1988.
J. Tierney, L. Wright, and K. Springen, "The Search for Adam and Eve," *Newsweek*, Jan. 11, 1988, pp. 46-52.
J. Wainscot, "Out of the Garden of Eden," *Nature* 325:13, 1986.
J.D. Watson, N.H. Hopkins, J.W. Roberts, et al., *Molecular Biology of the Gene* (Menlo Park, NJ: Benjamin/Cummings Publishing, 1987).

Box 3-F.—Molecular Anthropology

Anthropologists working in a central Florida bog recently discovered 8,000-year-old human skeletons with well-preserved brains, some of which have provided the oldest available samples of human DNA. Before this discovery, samples of DNA had been available only from the dried tissue remains of archaeological specimens from more arid regions. The fact that DNA can be preserved in other than excessively dry conditions greatly increases the number of archaeological sites at which more ancient DNA samples may be discovered. DNA fragments have also been prepared from Egyptian mummies, from an extinct animal called a quagga, and from a 35,000-year-old bison from Alaska. Biologists have been trying to clone these DNA fragments for use in studies of evolution. The sample from the extinct bison is likely to be old enough for comparison with modern buffalo DNA; this comparison may provide clues to the mechanisms of genome evolution. The human DNA samples, although important discoveries, are too recent to be particularly informative in studies of molecular evolution. As methods for working with the DNA extracted from these ancient species are improved, and as more specimens are uncovered, the application of gene mapping and sequencing technologies to anthropology and archaeology will be more feasible.

SOURCES:

B. Bower, "Human DNA Intact After 8,000 Years," *Science News*, Nov. 8, 1986, p. 293.

G.H. Doran, D.N. Dickel, W.E. Ballinger, et al., "Anatomical, Cellular and Molecular Analysis of 8,000-Year-Old Brain Tissue From the Windover Archaeological Site," *Nature* 323:803-806, 1986.

APPLICATIONS IN POPULATION BIOLOGY

Population biologists study populations by analyzing many individuals. They are interested in similarities and differences among individuals, among groups, among varieties, and among species. **To address such questions as how geography and environment affect inheritance patterns of certain traits, a physical map and a complete sequence of a single reference genome are not particularly valuable.** It would be more useful to have corresponding sequence information from widely diverse geographical areas, from various religious and ethnic subgroups, and from all races (9).

Population geneticists studying human beings, plants, or animals make great use of molecular markers—RFLPs and, increasingly, sequences of specific regions—to assess the extent of genetic variability (see box 3-G). Information on the same small chromosomal region (e.g., a gene or a region important for gene expression) from many individuals might be more useful than information on larger chromosomal regions from a few persons (43). Genes for rare diseases are not all found in a single human genome: Sickle cell hemoglobin, for instance, might not have been discovered if only Northern Europeans had been studied (4,9).

Problems in population genetics that bear on public health involve finding means for estimating human mutation rates,[1] for studying susceptibility to pathogens such as the virus responsible for acquired immune deficiency syndrome (AIDS), and for assessing possible environmental influences on these phenomena. The mechanisms generating physical variability among human beings are by no means well understood and involve not only genetic factors, but, among other things, a complex set of environmental factors. **DNA sequences from representative portions of many human genomes would also be of more immediate use than whole genome sequences for monitoring the effects of specific environmental factors on the structure of the human genome (9).**

---

[1] An OTA report assesses these scientific issues: U.S. Congress, Office of Technology Assessment, *Technologies for Detecting Heritable Mutations in Human Beings*, OTA-H-298 (Washington, DC: U.S. Government Printing Office, September 1986).

Box 3-G.—Implications of Genome Mapping for Agriculture

Since the dawn of agriculture, people have manipulated plants to enhance desired traits simply by observing the results of breeding, with no true understanding of the genetic principles involved. Many scientists working in the field of plant molecular biology believe that genome projects will have important implications for agriculture, by increasing knowledge about the genes that control or influence yield, time to maturation, nutritional content, resistance to disease, insects, and drought, and other factors in the production of crops.

The first gene maps ever constructed were assembled as a result of a series of painstakingly detailed crosses of pea plants and statistical analyses of data carried out by Austrian monk Gregor Mendel. Mendel was the first to recognize that some traits could be transmitted according to regular hereditary patterns. All modern genetics and much of modern biology build upon the foundation laid by Mendel.

Construction of RFLP marker maps has begun for corn, tomatoes, cabbage, and other crop plants. Such genetic maps give plant breeders the ability to use gene structure rather than observable characteristics to develop new varieties of plants. This ability should facilitate the development of intricate strategies for manipulating complex traits controlled by multiple, interacting genes.

The availability of RFLP maps makes it possible to select for several unrelated traits simultaneously or to manipulate traits controlled by clusters of genes that interact in complex ways. Researchers have mapped three genes that control efficiency of water use (drought tolerance) and five genes that have a major impact on flavor and soluble solids in tomatoes. Three genes that make a major contribution to insect resistance in tomatoes have also been mapped. A group of genes that influences yield has been found in corn. RFLP maps of genes influencing equally important traits are being developed for alfalfa, azaleas, cucumbers, onions, roses, sugar beets, and grasses.

Recently, there has been renewed interest in a small flowering plant called *Arabidopsis thaliana*, a duckweed in the mustard family. Although this plant has no obvious economic or nutritional value, it is a valuable research tool for plant molecular biologists. The *Arabidopsis* genome, at about 70 million base pairs, is about 10 percent or less the size of some of the major crop plant genomes, such as cotton, tobacco, or wheat. The small size of this plant makes it an important model system for studying general mechanisms of gene regulation that may be directly applicable to economically important but genetically less tractable plants. For these reasons, work has already begun on making complete genetic linkage and contig maps of the *Arabidopsis* genome.

SOURCES:
H. Bollinger, Native Plants Incorporated, personal communication, Jan. 19, 1988.
Judson, see app. A.
Mount, see app. A.
S.J. O'Brien, *Genetic Maps 1987* (Cold Spring Harbor, NY: Cold Spring Harbor Laboratory, 1987).
P.P. Pang, and E.M. Meyerowitz, "Arabidopsis Thaliana: A Model System for Plant Molecular Biology," *Biotechnology* 5:1177-1181, 1987.
M. Walton, and T. Helentjaris, "Application of Restriction Fragment Length Polymorphism (RFLP) Technology to Maize Breeding," unpublished manuscript, 1988.
J.D. Watson, N.H. Hopkins, J.W. Roberts, et al., in "Molecular Biology of the Gene," vol. I, (Menlo Park, NJ: Benjamin/Cummings Publishing, 1987).

CHAPTER 3 REFERENCES

1. Barker, D., Wright, E., Nguyen, K., et al., "Gene for von Recklinghausen Neurofibromatosis Is in the Pericentromeric Region of Chromosome 17," *Science* 236:1100-1102, 1987.
2. Bass, B.L., and Weintraub, H., "A Developmentally Regulated Activity That Unwinds RNA Duplexes," *Cell* 48:607-613, 1987.
3. Bodmer, W.F., Bailey, C.J., Bodmer, J., et al., "Localization of the Gene for Familial Adenomatous Polyposis in Chromosome 5," *Nature* 328:614-616, 1987.
4. Bowman, J.E., Department of Pathology, The University of Chicago, personal communication, December 1987.
5. Branden, C.-I., "Anatomy of a/b Proteins," *Current Communications in Molecular Biology: Computer Graphics and Molecular Modeling*, R. Fletterick and M. Zoller (eds.) (Cold Spring Harbor, NY: Cold

Spring Harbor Laboratory, 1986), pp. 45-51.

6. Cavenee, W.K., Hansen, M.F., Nordernskjoid, M., et al., "Genetic Origin of Mutations Predisposing to Retinoblastoma," *Science* 228:501-503, 1985.
7. Costs of Human Genome Projects, OTA workshop, Aug. 7, 1987.
8. Craik, C.S., Rutter, W.J., and Fletterick, R., "Splice Junctions: Association With Variation in Protein Structure," *Science* 204:264-271, 1983.
9. Crow, J.F., Laboratory of Genetics, University of Wisconsin, Madison, personal communication, May 1987.
10. Davies, K.E., Pearson, P.L., Harper, P.S., et al., "Linkage Analysis of Two Cloned Sequences Flanking the Duchenne Muscular Dystrophy Locus on the Short Arm of the Human X Chromosome," *Nucleic Acids Research* 11:2302-2312, 1983.
11. Donis-Keller, H., Green, P., Helms, C., et al., "A Genetic Linkage Map of the Human Genome," *Cell* 51:319-337, 1987.
12. Doolittle, R.F., *Of URFs and ORFs: A Primer on How To Analyze Derived Amino Acid Sequences* (Mill Valley, CA: University Science Books, 1987).
13. Doolittle, R.F., "Genes in Pieces: Were They Ever Together?" *Nature* 272:581-582, 1978.
14. Doolittle, R.F., Feng, D.F., Johnson, M.S., et al., "Relationships of Human Protein Sequences to Those of Other Organisms," *Molecular Biology of Homo Sapiens: Cold Spring Harbor Symposium on Quantitative Biology* 51(part 1):447-455, 1986.
15. Ecker, J.R., and Davis, R.W., "Inhibition of Gene Expression in Plant Cells by Expression of Antisense RNA," *Proceedings of the National Academy of Sciences USA* 83:5372-5376, 1986.
16. Egeland, J., Gerhard, D., Pauls, D., et al., "Bipolar Affective Disorders Linked to DNA Markers on Chromosome 11," *Nature* 325:783-787, 1987.
17. Estivill, X., Farrall, M., Scambler, P., et al., "A Candidate for the Cystic Fibrosis Locus Isolated by Selection for Methylation-Free Island," *Nature* 326:840-845, 1987.
18. Friend, S.H., Bernards, R., Rogelj, S., et al., "A Human DNA Segment With Properties of the Gene That Predisposes to Retinoblastoma and Osteosarcoma," *Nature* 323:643-646, 1987.
19. Gilbert, W., "Genes in Pieces Revisited," *Science* 228:823-824, 1985.
20. Gilbert, W., "Why Genes in Pieces?" *Nature* 271:501, 1978.
21. Gilbert, W., Marchionni, M., and McKnight, G., "On the Antiquity of Introns," *Cell* 46:151-154, 1986.
22. Go, M., "Correlation of DNA Exonic Regions With Protein Structural Units in Haemoglobin," *Nature* 271:90-92, 1981.
23. Gusella, J., Wexler, N., Conneally, P., et al., "A Polymorphic DNA Marker Genetically Linked to Huntington's Disease," *Nature* 306:234-238, 1983.
24. Herskowitz, I., "Functional Inactivation of Genes by Dominant Negative Mutations," *Nature* 329:219-222, 1987.
25. Knowlton, R.G., Cohen-Haguenauer, O., Van Cong, N., et al., "A Polymorphic DNA Marker Linked to Cystic Fibrosis Is Located on Chromosome 7," *Nature* 318:381-382, 1985.
26. Kucherlapti, R., "Gene Replacement by Homologous Recombination in Mammalian Cells," *Somatic Cell and Molecular Genetics* 13:447-449, 1987.
27. Lander, E.S., and Green, P., "Construction of Multilocus Genetic Linkage Maps in Humans," *Proceedings of the National Academy of Sciences USA* 84:2363-2367, 1987.
28. Lee, W.-H., Bookstein, R., Hong, F., et al., "Human Retinoblastoma Gene: Cloning, Identification, and Sequence," *Science* 235:1394-1399, 1987.
29. Mathew, C.G.P., Chin, K.S., Easton, D.F., et al., "A Linked Genetic Marker for Multiple Endocrine Neoplasia Type 2a on Chromosome 10," *Nature* 328:527-528, 1987.
30. McGarry, T.J., and Lindquist, S., "Inhibition of Heat Shock Protein Synthesis by Heat-Inducible Antisense RNA," *Proceedings of the National Academy of Sciences USA* 83:399-403, 1986.
31. McGinnis, W., Garber, R.L., Wirz, J., et al., "A Homologous Protein-Coding Sequence in *Drosophila* Homeotic Genes and Its Conservation in Other Metazoans," *Cell* 37:403-408, 1984.
32. Monaco, A., Neve, R., Colletti-Feener, C., et al., "Isolation of Candidate cDNAs for Portions of the Duchenne Muscular Dystrophy Gene," *Nature* 323:646-650, 1986.
33. National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
34. Nurse, P., and Lee, M.G., "Complementation Used To Clone a Human Homologue of the Fission Yeast Cell Cycle Control Gene *cdc2*," *Nature* 327:31-35, 1987.
35. Patrusky, B., "Homeoboxes: A Biological Rosetta Stone," *Mosaic* 18:26-35, 1987.
36. Rebagliati, M.R., and Melton, D.A., "Antisense RNA Injections in Fertilized Frog Eggs Reveal an RNA Duplex Unwinding Activity," *Cell* 48:599-605, 1987.
37. Reeders, S.T., Breuning, M.H., Davies, K.E., et al., "A Highly Polymorphic DNA Marker Linked to Adult Polycystic Kidney Disease on Chromosome 16," *Nature* 317:542-544, 1985.
38. Royer-Pokora, B., Kunkel, L.M., Monaco, A.P., et al., "Cloning the Gene for an Inherited Human

Disorder—Chronic Granulomatous Disease—on the Basis of Its Chromosomal Location," *Nature* 322:32-38, 1986.

39. Salmons, B., Groner, B., Friis, R. et al., "Expression of Antisense mRNA in h-ras Transfected NIH-3T3 Cells Does Not Suppress the Transformed Phenotype," *Gene* 45:215-220, 1986.
40. Scott, J., "Molecular Genetics of Common Diseases," *British Medical Journal* 295:769-771, 1987.
41. Seizinger, B.R., Rouleau, G.A., Ozelius, L.J., et al., "Genetic Linkage of von Recklinghausen Neurofibromatosis to the Nerve Growth Factor Receptor Gene," *Gene* 49:589-594, 1987.
42. Simpson, N.E., Kidd, K.K., Goodfellow, P.J., et al., "Assignment of Multiple Endocrine Neoplasia Type 2a to Chromosome 10 by Linkage," *Nature* 328:528-530, 1987.
43. Siniscalco, M., "On the Strategies and Priorities for Sequencing the Human Genome: A Personal View," *Trends In Genetics* 3:182-184, 1987.
44. Smithies, O., Gregg, R.G., Boggs, S.S., et al., *Nature* 317:230-234, 1985.
45. Stephens, J.C., Human Gene Mapping Library, Howard Hughes Medical Institute, New Haven, CT, personal communication, 1987.
46. St. George-Hyslop, P.H., Tanzi, R.E., and Polinsky, R.J., "The Genetic Defect Causing Familial Alzheimer's Disease Maps on Chromosome 21," *Science* 235:885-890, 1987.
47. Stein, J.P., Catterall, J.F., Kristo, P., et al., "Ovomucoid Intervening Sequences Specify Functional Domains and Generate Protein Polymorphism," *Cell* 21:681-687, 1980.
48. Stormo, G. 1987; "Identifying Coding Sequences," in *Nucleic Acid and Protein Sequence Analysis: A Practical Approach*, M.J. Bishop, and C.J. Rollings (eds.) (Oxford: IRL Press, 1987).
49. Sudhof, T.C., Russell, D.W., Goldstein, J.L., et al., "Cassette of Eight Exons Shared by Genes for LDL Receptor and EGF Precursor," *Science* 228:893-895, 1985.
50. Thomas, K.R., and Capecchi, M.R., *Nature* 324:34-38, 1986.
51. Thomas, K.R., Folger, K.R., and Capecchi, M.R., *Cell* 44:419-428, 1986.
52. Tsui, L.-C., Buchwald, M., Barker, D., et al., "Cystic Fibrosis Locus Defined by a Genetically Linked Polymorphic DNA Marker," *Science* 230:1054-1057, 1985.
53. U.S. Congress, Office of Technology Assessment, *New Developments in Biotechnology, 4: U.S. Investment in Biotechnology*, OTA-BA-360 (Washington, DC: U.S. Govenment Printing Office, in press).
54. U.S. Congress, Office of Technology Assessment, *Human Gene Therapy*, OTA-BP-BA-32 (Washington, DC: U.S. Government Printing Office, December 1984).
55. Wainwright, B.J., Scambler, P.J., Schmidtke, J., et al., "Localization of Cystic Fibrosis Locus to Human Chromosome 7cen-q22," *Nature* 318:384-386, 1985.
56. White, R., Woodward, S., Leppert, M., et al., "A Closely Linked Genetic Marker for Cystic Fibrosis," *Nature* 318:382-384, 1985.
57. Wise, J.A., Department of Biochemistry, University of Illinois, Urbana, personal communication, June 1987.

Chapter 4

Social and Ethical Considerations

CONTENTS

Page
Introduction 79
Basic Research 81
Levels of Resolution 81
Access and Ownership 82
Commercialization 82
Diagnostic/Therapeutic Gap 83
Physician Practice 83
Reproductive Choices 83
Eugenic Implications 84
  Positive Eugenics 84
  Negative Eugenics 85
  Eugenics of Normalcy 85
Attitudes 85
Role of Government 87
Duties Beyond Borders 87
Conclusion 88
Chapter 4 References 88

Boxes

Box *Page*
A. DNA Fingerprints 80
B. Determinism and the Human Genome 86

Chapter 4

Social and Ethical Considerations

"Science is a match that man has just gotta light. He thought he was in a room—in moments of devotion, a temple—and that his light would be reflected from and display walls inscribed with wonderful secrets and pillars carved with philosophical systems wrought into harmony. It is a curious sensation, now that the preliminary sputter is over and the flame burns up clear, to see his hands lit and just a glimpse of himself and the patch he stands on visible, and around him, in place of all that human comfort and beauty he anticipated—darkness still."

H.G. Wells, 1891

"The moral significance of humankind is no more threatened by peeking at the underlying musical notation, the base sequences, than is reading the score of Beethoven's last symphony diminishing to that piece of work."

Thomas H. Murray,
Case Western Reserve University, 1987

INTRODUCTION

As projects to map and sequence the human genome are undertaken, their long-range social and ethical implications need to be considered as part of policy analysis, yet further knowledge is needed before many of these implications emerge. Some will arise in the course of deciding what priority to give genome projects and what level of resolution (coarse genetic linkage map, complete DNA sequence) is most appropriate. More profound ethical questions are posed by possible applications of genetic data for altering the basis of human disease, human talents, and social behavior. Questions about personal freedom, privacy, and societal versus individual rights of access to genetic information are among the most important. A full picture of the human genome will of necessity raise questions about the desirability of using genetic information to control and shape the future of human society. The complexity and urgency of these issues will increase in proportion to advances in mapping and sequencing.

Part of the reason for studying genomes is to see how *variations* in genes account for differences among people. Some of the issues raised in this chapter relate specifically to these variations: What will be the impact of discovering that, in their genetic endowment, human beings are either more equal or more unequal than we now suppose? Other problems do not concern genetic differences, but rather the impact of discovering the extent to which genes do or do not limit the options of human beings in general. One commentator has argued that scientists bear a responsibility for using "moral imagination" to anticipate the full range of uses and consequences of their work, especially when that work is in the basic sciences (2).

The social considerations raised by genome projects include ethical issues. Ethical issues often arise in the context of debates about values, principles, or human actions that have had particular merit in the past. Such debates about what *ought* to be done often cannot be resolved by empirical inquiry. Specific genetic information such as the location of a gene along a chromosome or the sequence of nucleotide bases composing a specific gene is value-neutral and as such is not ethically troublesome. However, questions about private investment versus the allocation of Federal resources or about the proper use and availability of genetic information are ethical questions because they involve choices among actions based upon competing notions about what is good, right, or desirable.

Competing ideas about the desirable course of human action are developed from considerations about the greater good, personal freedom, benefiting others, avoiding harm, and fairness and equality. It is important to note that the ethical issues surrounding the use of and access to genetic information are not unique to the enterprise of mapping and sequencing the human genome (10) (see box 4-A). The existing uses of genetic screening, which in most cases are based on incomplete information about the location of a specific gene, already raise ethical questions. In addition, some general ethical questions are moot because of contemporary realities, for example, the question of whether there should be any human genome mapping and sequencing activities at all. This question is moot because mapping and sequencing projects have been underway for over a decade and

Box 4-A.—DNA Fingerprints

DNA fingerprints are derived from traces of human biological material such as blood, semen, hair, or other tissue. Recombinant DNA technology is applied to these samples to identify patterns of genetic sequence that are unique to each human being. Matched DNA fingerprints can establish the identity of a given individual with near certainty. DNA fingerprints, therefore, have great practical use in establishing the identity of criminals, family members, or bodily remains.

Genetic fingerprinting raises ethical issues such as the maintenance of personal autonomy when tissue samples are requested for identification purposes and the maintenance of confidentiality of individual genetic profiles. Even after tissue specimens have been discarded, there is considerable fear that genetic records will be retained in spite of the wishes of the human source of the tissue. California requires convicted sex offenders to give blood and saliva samples before their release from prison. The provision of such samples also makes it possible to discover information that may be incidental to past criminal records (e.g., XYY chromosome, drug use) but that could be used against the present or former inmates.

In the United States to date, practical applications of DNA fingerprinting have involved tests of specific suspects or known criminals. There are plans in California to store this information in the world's first computerized data bank of DNA fingerprints. In Great Britain, however, a DNA analysis of blood samples from all men and boys between the ages of 13 and 30 in Leicester County was conducted in an attempt to identify the person who raped and murdered two teenage girls. A 17-year-old boy originally charged with the crimes was released when his genetic profile did not match that derived from the semen left in the victims. More conventional investigative methods were later used to catch a suspect, a local baker who had avoided the test. The mass screening effort left investigators with a genetic profile on every young man in the county, information they later destroyed.

DNA fingerprinting has also been used as proof of paternity for immigration purposes. In 1986, Britain's Home Office received 12,000 immigration applications from the wives and children of Bangladeshi and Pakistani men residing in the United Kingdom. The burden of proof is on the applicant, but establishing the family identity can be difficult because of sketchy documentary evidence. Blood tests can also be inconclusive, but DNA fingerprinting results are accepted as proof of paternity by the Home Office.

Testing of extended families has been used in Argentina to identify the children of at least 9,000 Argentinians who disappeared between 1975 and 1983, abducted by special units of the ruling military and police. Many of the children born to the disappeared adults were kidnapped and adopted by military "parents," who claimed to be their biological parents. Once genetic testing of the extended family revealed the true identity of the child in question, the child was placed back in the home of its biological relatives. It was initially feared that transferring a child from its military "parents" who were kidnappers but who had nevertheless reared the child for years would be agonizing. In practice, the transferred children became integrated into their biological families with minimal trauma.

SOURCES:
Office of Technology Assessment, based on Herman, R., "British Police Embrace 'DNA' Fingerprints," *The Washington Post*, Nov. 24, 1987.
A.J. Jeffreys, J.F.Y. Brookfield, and R. Semeonoff, "Positive Identification of an Immigration Test-Case Using Human DNA Fingerprints," *Nature* 317:818-819, 1985.
J.M. Diamond, "Abducted Orphans Identified by Grandpaternity Testing," *Nature* 327:552-553 1987.

there has been no concerted effort to prohibit them. The more immediate questions, therefore, are how these projects should best proceed from now on and what use should be made of new genetic information.

Each of the following sections begins with a list of important social and ethical questions, followed by a short general discussion establishing the context of these issues and, in some cases, outlining opposing arguments. Decisions about mapping and sequencing rest in part on arguments about appropriate allocation of resources. Arguments about access to versus control of knowledge turn on debates about the relative importance of ethical principles such as autonomy (that is, self-determination or personal freedom of action) and beneficence (the duty to act in ways that benefit and do not inflict harm on others). There is general concern about the ways in which personal freedom of action might be either enhanced or diminished by increased knowledge about human genetics. Finally, there is significant concern about the possibility of *eugenics*, that is, that new and existing information will be used in attempts to improve hereditary qualities. The social and ethical arguments relevant to mapping and sequencing the human genome reveal the tension between an attempt to arrive at some clear insight about duties and obligations and an attempt to weigh benefits versus harms. The purpose of this chapter is to describe and clarify important points of social and ethical controversy, not to resolve them.

BASIC RESEARCH

- How should the conduct of research in the basic sciences, such as genome mapping and sequencing, be influenced by a concern for the social good?
- What are the considerations when basic research in the biological sciences seems to take resources away from areas of research that might have more immediate social benefit?

A genetic linkage map of the human genome already exists and progress has been made in the development of a physical map. Practical debate, therefore, centers on questions about the most efficient and effective way to develop the complete physical map, that is, whether the whole human genome should be sequenced in a systematic way and how new genetic information should be applied.

How these questions are answered depends upon the values attached to scientific progress and the relationship between scientific progress and human good. There is a strong argument that basic scientific research is valuable in and of itself and should be pursued for its own sake. Coordinated, systematic mapping of the human genome is consistent with this view, and proponents argue for resources and against constraints in the name of conducting *good science*. Others argue that scientists need to be responsive to and sometimes even constrained by the public interest (7).

LEVELS OF RESOLUTION

- What level of resolution of the physical map is really needed, and for what purposes?

While even a rough genetic map, permitting the identification of markers linked with major diseases, might prove useful to insurers or others bent on identifying high-risk individuals, it would have less value for basic researchers than a more precise map. From an ethical standpoint, the key arguments about levels of resolution, or molecular detail, are based on the distribution of costs and benefits involved. If the public is asked to pay an appreciable portion of the cost, then it deserves to participate in the political debate about embarking on an expensive, full-scale project. Scientific and technical factors being equal, chromosomal regions in which greater clarity would benefit many people (e.g., those associated with prevalent genetic diseases) might be addressed first. If the largest share of the costs is borne by the private sector, then few, if any, questions of priority will be posed, other than those chosen by the persons investing in the projects.

ACCESS AND OWNERSHIP

- What are the ethical considerations pertaining to control of knowledge and access to information generated by mapping and sequencing efforts?
- Who should have access to map and sequence information in data banks?
- Do scientists have a duty to share information; what are the practical extent and limits of such an obligation?
- Who owns genetic information?
- Do property rights to individuals' genetic identities adhere to them or to the human species (14)?
- Is genetic information merely a more detailed account of an individual's vital statistics, or should this information be treated as intrinsically private, not to be sought or disclosed without the individual's express consent (10)?

There is a method in scientific research that allows investigators to pursue their hunches, test their hypotheses, replicate their results, and publish their findings in roughly that order. Careful adherence to this process ensures accuracy and the orderly development of knowledge. The time lag between discovery of new information and communication of it, however, has caused some commentators to question whether scientists have the right to withhold information about genetic markers that might be of great interest to the public at large.

From an ethical perspective, it may be argued that genetic information is by definition in the public domain: The human genome is a collective property that should be held in common among all persons of human heritage (8). An opposing argument is that, since gene sequences are not commonly knowable and understanding them requires the use of expensive and often patentable machinery, discovery of sequences and the fruits that derive from them belong to the person who uncovered them. By this reasoning, it does not matter whether the sequences are unique or how they might be used, it is the labor and inventiveness associated with the discovery of them that makes them valid intellectual property. Current patent law takes the latter tack but limits patentability by preventing the patenting of a person or an idea.

One prominent scientist has acknowledged the public's special claim to the genome but argues that a public enterprise may not be the best way to satisfy this claim and that delay on so urgent a project serves no one (5). A significant portion of the value of the genetic information gathered through human genome projects will not be fully realized until some decades after the projects are completed, but there is little doubt that it will help elucidate the function and physical location of genes that cause or predispose to illness and disease. For this reason alone, the sequences will have substantial commercial value.

COMMERCIALIZATION

- What facets, if any, of human genome mapping and sequencing activities should be commercialized?

The commercial value of genome sequences has already been recognized by companies that have applied for patents on a number of specific materials and techniques. At least one company has argued that it has the right to copyright and control the materials and maps that it develops (5).

The selective forces of the marketplace have generated a database network, some portions of which are in the public domain and others of which are held by individual companies. The ethical issues of privatization of this knowledge turn on the importance of sequences lost to others by academic communities or corporations which have restricted the use of them. On one level, the problem is largely academic, since the data needed for a complete map and sequence could be assembled by the public sector, with duplication or purchase of the data held by private parties. On another level, however, the potential loss of critical data, the duplication of effort, and the control of knowledge raise serious questions about a combined scheme of public versus proprietary holding of fundamental knowledge. There is a strong argument that parts of research that are funded publicly should yield public information, while allowing scientists and others to retain the benefits of commercial exploitation of inventions.

DIAGNOSTIC/THERAPEUTIC GAP

- What are the ethical implications of the growing gap between diagnostic and therapeutic capabilities?
- Should diagnostic information about genetic disorders for which there is no therapeutic remedy be handled differently from that about disorders for which there *are* therapeutic interventions?

There is no doubt that continuing scientific advances in mapping and sequencing the human genome accelerate diagnostic applications. One philosopher has noted that the ability to map the human genome yields information about susceptibility that is more precise, more certain, and potentially more threatening to individual freedom and privacy than earlier methods of presymptomatic diagnosis and vague hypotheses about familial traits (10). A related issue is the need to protect information that may be available to or sought by third parties such as insurance companies or employers. Progress to date indicates that the ability to diagnose a genetic abnormality precedes the development of therapeutic interventions and that this gap may be growing. This is true for many genetic diseases, an important example being Huntington's disease (see box 7-A in ch. 7).

PHYSICIAN PRACTICE

- Do physicians and other health care providers face a conflict between an increasingly reductive approach to medical science and a focus on holistic patient care (17)?

Increased information about human genetics changes attitudes and alters the knowledge that serves as a basis for health care interventions. Physicians and other health care providers must constantly alter their views and understanding of human behavior, health, and disease. There are many examples of diseases that were once thought to be amenable to preventive health care that are now known to have a genetic component or cause. On a practical level this presents obvious difficulties, as health care providers are increasingly uncertain whether they are dealing with patterns of health and illness in individuals that can be ameliorated by changes in life style and medical treatment or if such patterns are in large part a matter of genetic destiny. In addition, the ethical principle of respect for persons indicates that individuals must be treated with care, compassion, and hope because they are persons and not merely the embodiments of a genetic formula or code.

REPRODUCTIVE CHOICES

- What ethical considerations arise from the increased ability of parents to determine the genetic endowment of their children (through such practices as selective termination of pregnancy, selective discarding of human embryos created in vitro, or selection of X- or Y-bearing sperm to determine the sex of the child)?

The ethical question of one generation's duties and obligations to another becomes more evident as genome mapping generates data pointing to the serious consequences of certain cultural practices or mating patterns. For example, it has been demonstrated that, if it were possible to choose the sex of their children, many individuals and couples would prefer that their firstborn be male

(18). It has also been demonstrated that firstborn children benefit from their early period of exclusive parental attention. If firstborn boys became the norm, it might further compromise equality of opportunity between men and women (16). In such circumstances, the conflicts among values and ethical principles such as autonomy, justice, and beneficence will be strong. Human mating that proceeds without the use of genetic data about the risks of transmitting diseases will produce greater mortality and medical costs than if carriers of potentially deleterious genes are alerted to their status and encouraged to mate with noncarriers or to use artificial insemination or other reproductive strategies (3).

On a practical level, the availability of information that couples might use to select embryos created in vitro has been hampered by an absence of federally funded research concerning many aspects of human fertilization. There has been a de facto moratorium on such research since 1980 (13).

EUGENIC IMPLICATIONS

- What ethical concerns arise from possible eugenic applications of mapping and sequencing data?

The possibility of mastery and control over human DNA once again raises the highly charged issue of genetic selection. One major difference between current and previous attempts at eugenic manipulation is that any potential eugenicist will have substantially more powerful techniques to effect desired ends and more data with which to muster support. With even the modest knowledge achieved in their first century, genetic techniques have become sophisticated enough to permit the use of selective breeding to produce animals with desired qualities.

When Francis Galton defined eugenics in 1883 as the science of improving the "stock," he intended the concept to extend to any techniques that might serve to increase the representation of those with "good genes." Thus, he indicated that eugenics was "by no means confined to questions of judicious mating, but takes cognisance of all the influences that tend, in however remote a degree, to give the more suitable races or strains of blood a better chance of prevailing speedily over the less suitable than they otherwise would have had" (4). Prior to the development of recombinant DNA technology, eugenic aims were primarily achieved by attempting to control social practices such as marriage. New technologies for identifying traits and altering genes make it possible for eugenic goals to be achieved through technological as opposed to social control.

Knowledge of human genetics will amplify the power to intervene in the diagnosis and treatment of disease. Each time a person who would otherwise have died of a disease caused or influenced by a gene is treated successfully by genetic or nongenetic means, the frequency of that gene in the population increases [Lappe, see app. A]. Human genome projects will intensify and accelerate the already difficult debates about who should have access to one's genetic information by providing faster and cheaper methods of testing for genetic variations, by making much more information available, and by increasing the specificity of genetic information (15). The ethical debate about eugenic applications more properly focuses on *how to use* new information rather than on *whether to discover* it. Eugenic programs are offensive because they single out particular people and therefore can be socially coercive and threatening to the ideas that human beings have dignity and are free agents.

Positive Eugenics

Beginning with Plato, philosophers have recognized that eugenic ends could be achieved through subtle or direct incentives to bring together presumptively fit human beings. Positive eugenics is defined here as the achievement of systematic or planned genetic changes in individuals or their offspring that improve overall human life and health and that can be achieved by programs that do not require direct manipulation of genetic material.

Most commentators have rejected or cast doubt on any uses of genetic engineering to enhance or directly improve the human condition. The President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research declared that efforts to improve or enhance normal people, as opposed to ameliorating the deleterious effects of genes, is at best problematic (11).

It may well be that the problem with positive eugenics has more to do with the means than with the ends. The basic objective of improving the human condition is generally supported, although debates about just what constitutes such improvement continue. Many concerns about eugenic policies in the past focused on the methods used to attain them, such as sterilization, rather than on the ends themselves.

Negative Eugenics

Negative eugenics refers to policies and programs that are intended to reduce the occurrence of genetically determined disease. It implies the selective elimination of gametes (ova or sperm) and fetuses that carry deleterious genes, as well as the discouraging of carriers of markers for genetic disease from procreation. There are few technical obstacles to karyotyping human beings for eugenic reasons. Verbal genetic histories of sperm donors, for example, are designed to exclude donors carrying some genetic diseases. Such a screening process, accompanied by a physical examination and laboratory tests, has already been recommended by the Ethics Committee of the American Fertility Society (1). The development of specific genetic tests could make gamete screening easier and more specific and will also expand existing capabilities to conduct prenatal tests.

Eugenics of Normalcy

The third eugenic use of genetic information would be to ensure not merely that a person lacks severe incapacitating genetic conditions, but that each individual has at least a modicum of normal genes. One commentator has argued that individuals have a paramount right to be born with a normal, adequate hereditary endowment (6). This argument is based on the idea that there can be some consensus about the nature of a normal genetic endowment for different groups of the human species. The idea of genetic normalcy, once far-fetched, is drawing closer with the development of a full genetic map and sequence; however, concepts of what is normal will always be influenced by cultural variations and subject to considerable debate.

ATTITUDES

- How will a complete map and sequence of the human genome transform attitudes and perceptions of ourselves and others?

One of the strongest arguments for supporting human genome projects is that they will provide knowledge about the determinants of the human condition. One group of scientists has urged support of human genome projects because sequencing the human genome will provide one of the most powerful tools humankind has ever had for deciphering the mysteries of its own existence (12).

The relevance of this proposition will depend on the degree to which complex human behaviors are determined by understandable genetic factors. It will also depend on how important human genome projects are to understanding genetic factors for complex traits. Whether higher human attributes are reducible to molecular constructs is a topic of considerable debate in the philosophy of biology, and human genome projects would doubtless enlarge and intensify this debate. A reasonable hypothesis is that, while little information of direct or immediate value regarding complex behaviors is likely to result from human genome projects, insights into the possible construction of control regions for the development of the human embryo, the genetic basis for organizing neuronal pathways, and the genetic control of sexual differentiation will all be significantly enhanced. In the long run, knowledge of human genetics will make scientific understanding of human life more sophisticated.

A greatly increased understanding of how genes shape characteristics could influence human beings' attitudes toward themselves and others [Glover, see app. A]. Such increased understanding might highlight the degree to which genetic factors are equal or unequal for traits that confer social advantage. This information might reveal that human beings have fewer options than they suppose and could thereby encourage a determinist view of human choices (see box 4-B), or it could reveal just the opposite. A general increase in genetic information might also alter social customs based on erroneous scientific assumptions.

Many individuals have general beliefs about their genetic potential for achievement in certain spheres of activity, about the limits of possible improvement through effort or environmental change. These intuitive beliefs are often vague and inaccurate. Often, it is only in regard to a few skills or characteristics that individuals have pushed against the limits of their potential. When science makes it possible to trace the actual limits of individuals, intuitive perceptions may turn out to be wrong. This has the potential of both enhancing and limiting personal liberty.

Box 4-B.—Determinism and the Human Genome

Determinism in biology is the general thesis that, for every action taken, there are causal mechanisms that preclude any other action. Mapping and sequencing the human genome will not alone impose a determinist view of human nature. Seeing where genes are located, or knowing the order of bases in the DNA, will not alone make behavior predictable.

But mapping and sequencing *together with* tracing the pathways between genes and behavior will start to paint a determinist picture. Scientists are now starting to work out these pathways. Take, for example, the pattern of behavior classified by psychiatrists as sensation seeking, which involves a disposition toward gambling and alcoholism. This behavior is correlated with low levels of activity of the platelet monoamine oxidase. These levels of activity have been shown by studies of twins to be largely under genetic control.

In a determinist model, human actions can be explained in terms of causal mechanisms, even though those mechanisms may be very complex. If this model is right, it seems that what human beings do, just as much as what billiard balls do, is the product of a set of laws operating in particular circumstances.

This view of human nature is disturbing. It suggests that a Godlike scientist, with complete knowledge of all the relevant causal laws and of the circumstances in which they operate, could successfully predict human action. In two different ways, determinism is at least an apparent threat to our attitudes. First, the elimination of genuine choice would leave no room for the belief that we can partly create, actualize, or modify ourselves. Second, undermining choice may also undermine many emotional reactions to others. The determinist picture may not leave room for justifiable resentment of what people do or for justifiable feelings of blame or guilt.

There are alternative views within determinism. *Hard determinism* is the view that individual choice is entirely ruled out, along with the emotional responses linked to holding people responsible for what they do. *Soft determinism* asserts that free choice and responsibility are compatible with determinism.

The issue is whether the soft determinist can resist the hard determinist's argument against freedom and the reactive attitudes. There are two strategies for resisting: 1) to point out that determinism is not the same as fatalism, that even in a deterministic world what human beings do influences the future; and 2) to disagree that determinism eliminates genuine choice, attempting to work out a model of free action that is compatible with determinism.

SOURCES:
Office of Technology Assessment, 1988.
Glover, see app. A.

ROLE OF GOVERNMENT

- What is the proper role of government in mapping and sequencing the human genome?
- Specifically, does the government have a role in deciding what data should be collected in gene mapping and sequencing? How should this information be disseminated and guarded from abuse?

The lines of power, coercion, and authority in the public and private scientific sectors are blurred because the first genetic maps are being made in corporations (e.g., Collaborative Research, Inc.) and in private philanthropies based in universities (e.g., the Howard Hughes Medical Institute at the University of Utah).

The ethical arguments for involving the Federal Government in the process of genome mapping, whether by shaping, constraining, blocking, or doing nothing, center on the public interest in making resources available in ways that are consistent with the considerations of beneficence, justice, and autonomy. These issues encompass academic freedom or freedom of scientific inquiry because the projects have universal and lasting implications. Once the human genome is mapped and sequenced, the resulting data will have widespread implications for generations to come [Lappe, see app. A].

The precise boundary between basic and applied science is hard to draw, but there is enough understanding of where it lies to be able to use it as a basis for policy. A case might very well be made for a government policy that would leave basic research unrestricted but that would place some stringent controls on applied research and technological applications, for example, by ensuring that genetic testing is voluntary and access to data is controlled.

All research carries with it the likelihood of changing one's conception of the world and so of changing one's attitudes. For these reasons, there is a strong case against government intervention to stop research. There are four main arguments:

1. Stopping research might be opting for comfortable ignorance or illusion rather than uncomfortable truth. The growth of science has rested on the preference for uncomfortable truth. Those who view science as one of mankind's finest creations will be dismayed at any wholesale repudiation of this preference.
2. It is unlikely that existing world views, beliefs, and attitudes can be protected by shutting down basic research. The knowledge that such protection was needed might itself start to undermine existing views.
3. As a practical matter, it may be that government cannot stop basic research. It is not easy to monitor what goes on in laboratories, and what is stopped in one country may take place in another [Glover, see app. A].
4. Stopping research blocks both possible benefits and risks. The belief that research can be performed to permit benefits while coping with and occasionally avoiding risks is a matter of historical precedent.

DUTIES BEYOND BORDERS

- What, if any, ethical issues are raised when considerations of international competitiveness influence basic scientific research?
- What, if any, are the duties and obligations of the United States to disseminate mapping and sequencing information abroad?
- What are the implications of shared information for international competitiveness?
- What are the international implications of sharing technological applications of mapping and sequencing information?
- What issues are involved when applications of genetic information or biotechnology that are of great use to Third World countries are not developed or fully exploited because they are less profitable for industrialized countries?

The United States has recently proposed an international framework of rules for science. The purpose of this framework is to see that all nations do their fair share of basic research and that all the results of such research be made public, except for those with strategic implications (9). The increased protection of intellectual property and patent rights for technological innovations formed the basis of this proposal; these rights were also central to recent international trade talks. There is some sentiment that barriers to the transfer of technology would continue even if there were no reward for intellectual property. One commentator has noted that, unless products are protected by a set of principles now, basic scientific results could become increasingly restricted; some nations might do less basic research and instead emphasize applying other nations' results (9).

The most common single-gene defects, disorders of the hemoglobin molecules that carry oxygen in red blood cells, are highly prevalent in many nations in Southern Europe, Africa, the Middle East, and Asia. Such nations would benefit most if research tools became widely available as they were developed and if priorities for which chromosomal regions are mapped first took world prevalence of disorders into account. Use of map and sequence information by developing nations may also require special attention to devising screening tests that are cheap and simple, and might entail access to services (e.g., sequencing or mapping) located in developed nations [Weatherall, see app. A].

CONCLUSION

All human beings have a vital interest in the social and ethical implications of mapping and sequencing the human genome. It is not surprising, therefore, that there are debates about how genome projects should proceed. These extend beyond considerations of scientific efficacy and involve the interests of patients, research subjects, physicians, academicians, lawyers, entrepreneurs, and politicians. Mapping the human genome accelerates our rate of understanding—and the distance between increased understanding and direct intervention to alter the human genome is shrinking. Add to this the development of scientific tools such as gene probes, and immediate practical questions are posed: How should basic research be conducted? What level of resolution in mapping is necessary? Who should have access to and ownership of data banks and clone repositories? How should thorny questions surrounding commercialization be handled? Long-range questions about eugenics, reproductive choices, the role of government, and possible duties and obligations beyond national borders also arise. These questions are complex and are not likely to be resolved in the near future. It will therefore be necessary to ensure that some means for explicitly addressing ethical issues attends scientific work.

CHAPTER 4 REFERENCES

1. American Fertility Society, Ethics Committee, "Ethical Considerations of the New Reproductive Technologies," *Fertility and Sterility* 46:1S–94S, 1986.
2. Callahan, D., "Ethical Responsibility in Science in the Face of Uncertain Consequences," *Annals of the New York Academy of Sciences* 265:1-12, 1976.
3. Campbell, R.B., "The Effects of Genetic Screening and Assortative Mating on Lethal Recessive-Allele Frequencies and Homozygote Incidence," *American Journal of Human Genetics* 41:671-677, 1987.
4. Galton, F., *Inquiries Into Human Faculty* (London: Macmillan, 1883).
5. Gilbert, W., quoted in Roberts, L., "Who Owns the Human Genome?" *Science* 237:358–361, 1987.
6. Glass, B., "Ethical Problems Raised by Genetics," in *Genetics and the Quality of Life*, Charles Birch and Paul Abrecht (eds.) (New York, NY: Pergamon Press, 1974), pp. 51-57.
7. Institute of Medicine, *Responding to Health Needs and Scientific Opportunity: The Organizational Structure of the National Institutes of Health* (Washington, DC: National Academy Press, 1984).
8. Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
9. MacKenzie, D., "US Advocates a 'World Code' for Science," *New Scientist* 5 (November): 245, 1987.
10. Macklin, R., "Mapping the Human Genome: Problems of Privacy and Free Choice," *Genetics and the*

Law III, A. Milunsky and G.J. Annas (eds.) (New York, NY: Plenum Press, 1985), pp. 107-114.

11. President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, *Screening and Counseling for Genetic Conditions: A Report on the Ethical, Social, and Legal Implications for Genetic Screening, Counseling and Education Programs* (Washington, DC: U.S. Government Printing Office, 1983).
12. Smith, L., and Hood, L., "Mapping and Sequencing the Human Genome: How To Proceed," *BioTechnology* 5:933-939, 1987.
13. U.S. Congress, Office of Technology Assessment, *Infertility: Medical and Social Choices*, OTA-BA-358 (Washington, DC: U.S. Government Printing Office, May 1988).
14. U.S. Congress, Office of Technology Assessment, *New Developments in Biotechnology: Ownership of Human Tissues and Cells*, OTA-BA-337 (Washington, DC: U.S. Government Printing Office, March 1987).
15. U.S. Congress, Office of Technology Assessment, *Human Gene Therapy*, OTA-BP-BA-32 (Washington, DC: U.S. Government Printing Office, December, 1984), app. B.
16. Walters, L., "Is Sex Selection a Frivolous Practice? Yes," *Washington Post*, Apr. 7, 1987, p. 11.
17. Weatherall, D., "Molecules and Man," *Nature* 328: 771-772, 1987.
18. Westoff, C.F., "Sex Preselection in the United States, Some Implications," *Science* 184:633-636, 1974.

Chapter 5

Agencies and Organizations in the United States

CONTENTS

Page
National Institutes of Health 93
  The Institutes 95
  Funding Mechanisms 95
  Research Infrastructure 96
  Peer Review 98
Department of Energy 99
  The Office of Health and Environmental Research 99
  Health and Environmental Research Advisory Committee Report 101
National Science Foundation 102
National Bureau of Standards 103
Centers for Disease Control 103
Department of Defense 104
Office of Science and Technology Policy 104
Domestic Policy Council 105
Office of Management and Budget 105
Howard Hughes Medical Institute 105
National Research Council 106
Private Corporations 107
Private Foundations 109
Summary 109
Chapter 5 References 110

Figures

Figure *Page*
5-1. Organization of NIH 94
5-2. Organization of DOE 99
5-3. Howard Hughes Medical Institute Genomics Resources 106

Tables

Table *Page*
5-1. NIH Support for Mapping and Sequencing, Fiscal Year 1986 94
5-2. Division of Research Resources Activities Related to Molecular Genetics 97
5-3. Budget for Proposed DOE Human Genome Initiative 101

Chapter 5

Agencies and Organizations in the United States

> "Science has been a formative factor in making both the Federal Government and the American mind what they are today. The relation of the government to science has been a meeting point of American political practice and the nation's intellectual life."
>
> A. Hunter Dupree,
> *Science and the Federal Government*
> (Baltimore: Johns Hopkins University Press, 1986), p. 2.

Projects involving or related to mapping or sequencing the human genome can be found in several Federal agencies and nongovernment organizations in the United States. Activities at the principal agencies and organizations will be briefly reviewed in this chapter. They include:

- Federal agencies:
  —the National Institutes of Health,
  —Department of Energy,
  —the National Science Foundation,
  —the National Bureau of Standards,
  —the Centers for Disease Control,
  —the Department of Defense,
  —the Office of Science and Technology Policy,
  —the Office of Management and Budget,
  —the Domestic Policy Council;
- Nongovernment organizations:
  —the Howard Hughes Medical Institute,
  —the National Research Council,
  —private corporations,
  —private biomedical research foundations and other philanthropies.

NATIONAL INSTITUTES OF HEALTH

The National Institutes of Health (NIH) form one branch of the Public Health Service of the U.S. Department of Health and Human Services. They are administered by a director, currently James Wyngaarden. NIH is a highly decentralized confederation of institutes, divisions, bureaus, and the National Library of Medicine (see figure 5-1). The principal mission of NIH is to conduct and support biomedical research to improve human health.

NIH was established in 1887. Since World War II, it has become "the foremost biomedical research facility not only in the United States but in the world" and "a brilliant jewel in the crown" of the Federal Government, according to Wilbur Cohen, former Secretary of the Department of Health, Education, and Welfare (10).

The institutes with the largest budgets for genetic mapping and DNA sequencing are the National Institute of General Medical Sciences, the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, the National Institute of Child Health and Human Development, and the National Institute of Neurological and Communicative Disorders and Stroke (see table 5-1). In fiscal year 1986, NIH supported approximately 3,000 projects that involved mapping or sequencing, with a combined budget of $294 million (out of a total budget of $5.26 billion, of which all but 5 percent went for research activities) (18). NIH estimated it spent $313 million for such projects in fiscal year 1987 (18).

Planning at NIH is decentralized. The Office of the Director has responsibility for overall direction, but most programmatic decisions are made in the institutes, which are autonomous and largely control their own budgets. A 1984 report by the Institute of Medicine remarked on the "absence of the trappings of bureaucratic authority; hence the Director manages largely on the basis of persuasion, consensus, and knowledge" (11).

Figure 5-1.—Organization of NIH

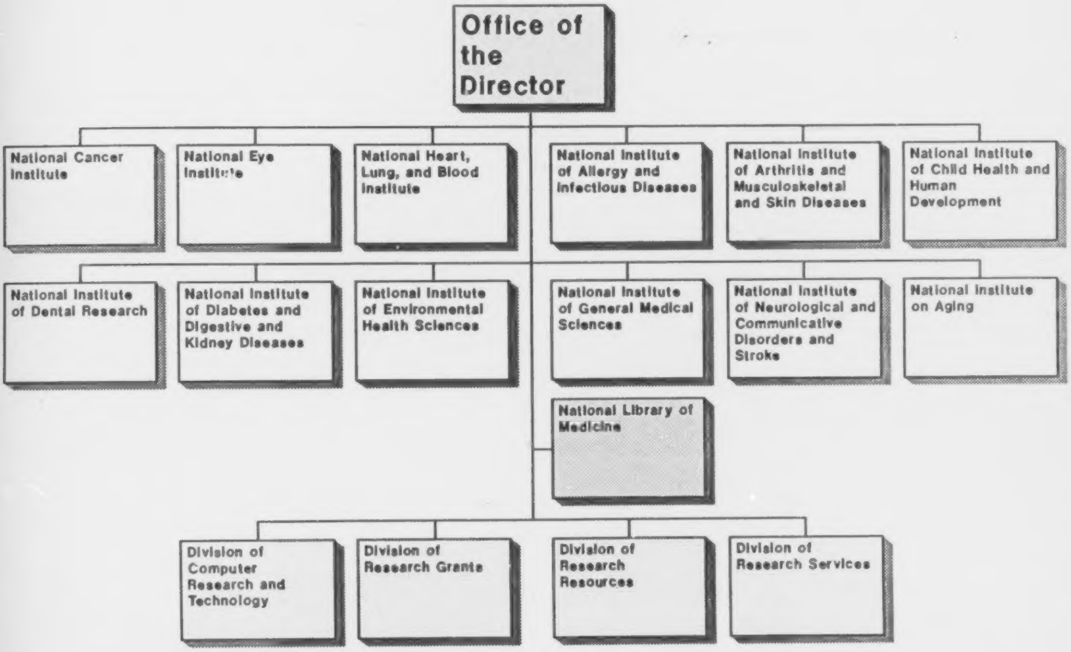

Office of the Director
National Cancer Institute
National Eye Institute
National Heart, Lung, and Blood Institute
National Institute of Allergy and Infectious Diseases
National Institute of Arthritis and Musculoskeletal and Skin Diseases
National Institute of Child Health and Human Development
National Institute of Dental Research
National Institute of Diabetes and Digestive and Kidney Diseases
National Institute of Environmental Health Sciences
National Institute of General Medical Sciences
National Institute of Neurological and Communicative Disorders and Stroke
National Institute on Aging
National Library of Medicine
Division of Computer Research and Technology
Division of Research Grants
Division of Research Resources
Division of Research Services

Table 5-1.—NIH Support for Mapping and Sequencing, Fiscal Year 1986 (millions of dollars)

Institute	Human research	Nonhuman research	Total
NIGMS	12.4	99.6	112.0
NCI	18.3	24.2	42.5
NIAID	6.0	28.0	34.0
NICHD	11.8	18.2	30.0
NINCDS	10.7	10.6	21.3
Other institutes and divisions and the NLM	31.9	22.1	54.0
Total	91.1	203.0	294.0

Abbreviations: NIGMS = Institute of General Medical Sciences; NCI = National Cancer Institute; NIAID = National Institute of Allergy and Infectious Diseases; NICHD = National Institute of Child Health and Human Development; NINCDS = National Institute of Neurological and Communicative Disorders and Stroke; NLM = Library of Medicine.

SOURCE: Office of the Director, National Institutes of Health, May 1987, as modified by the Office of Technology Assessment.

Coordination of the various institutes is accomplished largely by the Office of the Director. In October 1986, the Advisory Committee to the Director held a meeting at NIH entitled "The Human Genome," at which views about setting NIH policy were presented by many experts. Statements both in favor of and against special initiatives were aired (22). A working group of NIH administrators was formed subsequent to that meeting. The working group is chaired by the director; other members represent several of the institutes and divisions most directly involved.[1] This working group is responsible for setting overall policies for NIH in connection with human genome projects, and it initiated two program announcements in 1987 (17). Included in NIH's related research figures are several grants to produce physical maps of other organisms or parts of human chromosomes, to develop cloning or DNA detection techniques, and to develop other relevant technologies. The new genome programs are to develop new methods for analysis of complex genomes and to improve computer representation and analysis of information derived from molecular biology (12). These solicitations for proposals were not associated with any new or additional funding in 1987, but Congress has set aside $17.2 million for them in 1988. The budget request for fiscal year 1989 is $28 million. To review complex genome and informatics proposals, NIH will convene new peer review committees.

---

[1] Other members of the NIH working group on the human genome are Ruth Kirschstein (Director of the National Institute of General Medical Sciences), Betty Pickett (Director of the Division of Research Resources), Duane Alexander (Director of the National Institute of Child Health and Human Development), Donald A.B. Lindberg (Director of the National Library of Medicine), Jay Moskowitz (Associate Director for Program Planning and Evaluation), and George Palade (Yale University). Rachel Levinson (Office of the Director) is executive secretary.

The NIH also plans to seek advice from outside scientists and to keep congressional staff abreast of genome projects through a series of workshops, the first of which was held February 29 and March 1, 1988.

The Institutes

The National Institute of General Medical Sciences (NIGMS) supports research and training in the basic biomedical sciences fundamental to understanding health and disease. Its primary function is to support research projects conducted by scientists throughout the nation and the world that can serve as the bases for the more disease-oriented research undertaken by the other NIH institutes. NIGMS will administer the funds set aside for characterization of complex genomes. Unlike most of the other NIH institutes, NIGMS has no intramural research program—its funding is for work done by non-NIH scientists.

NIGMS supports a major share of basic research in genetics, including research on nonhuman species. Such work is concentrated at NIGMS because the institute is responsible for research related to fundamental biology or a broad array of disorders rather than to a disease group, developmental stage, or organ system. Genetics underlies many physiological processes and can explain many disease states, but most fundamental genetics research is not designed to elucidate a single disease; rather, it elucidates general mechanisms or illuminates how human diseases might occur by showing how other organisms function. Understanding other organisms is often the first and most important step in understanding human health and disease, but the details of how knowledge about bacteria, yeast, or animals will relate to human biology is rarely known in advance. These are some of the reasons that NIGMS supports such a large share of the work on genetics of nonhuman organisms.

Each NIH institute other than NIGMS has as its mission the support of research on a range of diseases. The range of diseases may be defined by organ system, developmental stage, explicitly named disease group, or other criteria (see figure 5-1). The distinction between the kinds of work supported by NIGMS and by the other institutes is not hard and fast; in fact, support extends over a broad range of scientific projects that could come under the aegis of NIGMS or one of the other institutes. The National Institute of Child Health and Human Development (NICHD), for example, has a program that investigates the basic molecular biology of development. In connection with this, NICHD convened in May 1987 a meeting of scientists working on human chromosome 21. Chromosome 21 is of special interest to persons doing research on Down's syndrome, Alzheimer's disease, and several other diseases; it is also of interest because it contains the genes underlying several important and well-characterized biochemical processes.

All of the institutes support genetic research (in fact, other institutes support more of it in the aggregate than NIGMS), but this research is often directed at finding the location of a particular disease-associated gene. (For example, study of the familial form of Alzheimer's disease is supported by the National Institute of Neurological and Communicative Disorders and Stroke, the National Institute on Aging, and the National Institute on Mental Health.) Institutes develop preventive diagnostic tests and therapies for genetic diseases.

Funding Mechanisms

Spending at NIH is predominantly for investigator-initiated, basic, undirected research. Most projects are related to human diseases, animal models of human disease, or fundamental research on biological questions that might illuminate human biology in health and disease. NIH's primary funding mechanism is the investigator-initiated scientific grant (classified as R01 by the NIH bureaucracy and widely known by that term) awarded to a single investigator or small group. The typical R01 grant (and there are now more than 6,100 of them) is given to a research scientist at a university or other research center in response to a proposal submitted by that scientist. The proposal outlines the research question addressed, the approach to the question, the people who would work on the project, and the budget for the project. The average grant amount for projects that involved mapping or sequencing was $130,000 in 1986 (5).

Some efforts—those with a specific purpose—are more amenable to funding by contract. The GenBank® database of nucleic acid sequences, for example, is supported by this mechanism under a $17.2 million 5-year contract with Intelligenetics Corp. of Mountain View, California (with a subcontract to the Los Alamos National Laboratory, where GenBank® is housed). NIGMS administers the GenBank® contract and is the principal funding unit, with contributions from other NIH institutes and divisions, the Department of Energy, and the National Science Foundation. NIGMS also maintains the Human Mutant Cell Repository under contract. This is a resource for persons attempting to use genetic techniques to understand diseases or physiological processes. In these instances and others, NIH can contract with a provider to deliver a service.

Each NIH institute other than NIGMS administers a program of intramural research, in most cases located on the NIH campus. Investigators are employed directly by NIH. The intramural research programs of NIH collectively constitute the largest biomedical research facility in the world. NIH's intramural research complements its extramural support of university and research center scientists. Components of human genome projects that require direct management by NIH or that would be best integrated into existing programs could be added to the intramural research programs.

NIH is not often associated with large, centrally administered programs, but it does support many. The National Cancer Institute, for example, supports a number of centers that bring research, training, information dissemination, and clinical application under one roof or administrative arrangement in order to accelerate the communication of ideas among normally disparate groups. Program and center grants are typically larger than investigator-initiated grants and can include funds for training as well as for equipment and research materials. A concerted research program has recently begun to combat AIDS (acquired immunodeficiency syndrome). NIH has the capacity to direct a research program that requires coordination and some central planning.

Research Infrastructure

A small but important fraction of NIH funding goes to support a research infrastructure—resources used by a wide array of scientists and clinical investigators to facilitate their research. Much of the support for a research infrastructure comes from the Division of Research Resources (DRR) at NIH. Databases for genetic information, funding for repositories (e.g., for human cell lines, DNA clones, and probes), and support of the National Library of Medicine are also important components of the research infrastructure for mapping and sequencing.

The DRR supports regional and national centers with various purposes. It is divided into five programs, several of which support projects relevant to mapping and sequencing. One of the purposes of DRR-supported resources is to provide scientists and clinicians with access to advanced research technologies. This involves support of several databases, materials repositories, computer resource centers, and grants to generate and analyze biomedical research data (see table 5-2). DRR cofunds with NICHD the Repository of Human DNA Probes and Libraries. This repository facilitates exchange of research materials crucial to genome projects. DRR also supports a grants program to apply artificial intelligence and other sophisticated approaches of information science to understanding sequence data and managing large masses of biological information. Several databases, repositories, and activities supported by DRR are cofunded by other NIH institutes or agencies of the Federal Government. DRR funds its resources through grants and contracts, primarily to nongovernment scientists. It has helped fund two workshops directly related to human genome

Table 5-2.—Division of Research Resources Activities Related to Molecular Genetics

Resource	Function	Location
Protein Identification Resource	Database for protein sequences and software	Georgetown University Washington, DC
Sequences of Proteins of Immunological Interest	Annotated protein sequence file	Bolt, Beranek, & Newman, Inc. Boston, MA
BIONET	Network, database linkage, and software for use in molecular biology	Intelligenetics Mountain View, CA
Dana Farber Cancer Institute and Baylor College of Medicine (with other NIH institutes)	DNA sequence analysis software and other computer resources	Boston, MA and Houston, TX, respectively
National Flow Cytometry Resource (with DOE, other NIH Institutes)	Chromosome and cell sorting	Los Alamos National Laboratory Los Alamos, NM
DNA Segment Library (with NICHD, DOE)	Distribution center for cloned human DNA made by Los Alamos and Lawrence Livermore national laboratories	American Type Culture Collection Rockville, MD
Cell Line Two-Dimensional Gel Electrophoresis Database	Cell line analysis by protein electrophoresis in two dimensions	Cold Spring Harbor Laboratory Cold Spring, NY

SOURCE: Office of Technology Assessment, 1988.

projects—one with the Department of Energy (DOE) on materials repositories and databases, the other with DOE, the National Library of Medicine, and the Sloan Foundation on applying information management systems to analysis of complex biological problems.

The National Library of Medicine (NLM) is the largest and most comprehensive collection of medical information in the world. The library also supports an extensive medical bibliographic resource —the published *Index Medicus* and MEDLARS/MEDLINE, the most widely used on-line computer reference service for medicine and biomedical research. The NLM has been called "the foremost biomedical communications center in the world" (10) and the "central nervous system of American medical thought and research" (16).

The library was started in 1818 as a few books in the office of the Surgeon General of the Army, Joseph Lovell. Its great flowering occurred under John Shaw Billings in the period after the Civil War, when it became an internationally recognized medical library. The library was transferred from the military to the civilian sector in 1956, and its name was changed to the National Library of Medicine through legislation sponsored by Senators Lister Hill and John Kennedy. A new building for the collection was constructed on the NIH campus in 1962, and in 1980 the 10-story Lister Hill Center was dedicated. The library became part of NIH in 1968 (16).

The NLM's expertise lies in managing clinical and biomedical research information. This includes not only storage of books and journals, but the publication of reference works that list the extensive international biomedical literature and the maintenance of computer databases that make access to the medical information more efficient. In recent years, the Board of Regents of the NLM has pointed to biotechnology databases as an area of expected future growth and has encouraged library staff to provide improved access to databases relevant to genetics, molecular biology, and other aspects of the "new biology."

Late in the 99th Congress, Senator Claude Pepper introduced a bill, the National Center for Biotechnology Information Act of 1986, that would give NLM responsibility to "develop new communications tools and serve as a repository and as a center for the distribution of molecular biology information" (H.R. 99-5271). The bill was reintroduced early in the 100th Congress with minor modifications (H.R. 100-393), and a companion measure with very similar provisions (S. 100-1354) was introduced in the Senate by Lawton Chiles. The bill was further amended and introduced as

S. 100-1966 jointly by Senators Chiles, Kennedy, Domenici, Leahy, Graham, and Wilson in December 1987. These bills would make the NLM responsible for improving access to the numerous databases used in molecular biology and clinical genetics, with funding authorized at $10 million per year for fiscal years 1988 through 1992. Appropriations for fiscal year 1988 included $3.83 million for these purposes (13).

The NLM has been conducting research on how to make human genetic information available to the medical community for several years. It has made Victor McKusick's *Mendelian Inheritance in Man* (15), the pivotal catalog of human genetic loci identified by analysis of pedigrees, available on-line through its Information Retrieval Experiment program, and it has linked the data in this volume to information available in GenBank® and the Protein Identification Resource databank. The library has also begun an experimental program to link molecular biology databases, using researchers on the NIH campus in Bethesda to test the system. It plans to make DNA sequence and protein database analysis possible through a computer link to the National Cancer Institute's supercomputer center in Frederick, Maryland. The Howard Hughes Medical Institute and the NLM have been discussing ways to link access to the various databases supported by NIH and the institute.

Peer Review

The National Cancer Institute became the first American institution to routinely employ peer review when it established the National Cancer Advisory Council in 1937 (26). Since then, peer review has become an essential element in allocating funds for research grants at NIH. The review system is two-tiered: The initial review is done by study sections of scientific experts; the second tier involves recommendations for funding made by an institute advisory council.

Review of the typical grant involves several steps. A grant application is received by the Division of Research Grants at NIH from an investigator (or from a program or center) under sponsorship of an institution. The application is then assigned to a group of scientists from a particular discipline appointed by the Director of NIH. These groups meet three times a year to review grant applications. They assess applications for their scientific merit (including originality, feasibility, and importance), the competence of the investigators to do the work, and the appropriateness of the proposed budget (4). The study section votes to approve, disapprove, or defer consideration of an application. For grant applications that are not defended, a priority score ranging from 100 (best) to 500 is assigned, based on the rankings of the individual members of the study section. This priority score is then included in a summary statement for each application that briefly states reviewers' opinions. The summary (and where necessary the full documentation) is then passed on to the appropriate advisory council.

Each institute at NIH has an advisory council, composed of eminent scientists and informed lay members, that recommends applications for funding. Advisory councils monitor the quality and fairness of review by the study sections and assess special relevance to important national health needs and the mission of the institute. Members of advisory councils are appointed by the Secretary of Health and Human Services, except those on the National Cancer Advisory Board, who are appointed by the President. In most cases, the advisory council approves the actions of the study sections. Fewer than 10 percent of grant applications are singled out for special discussion or action by the advisory councils (26).

Staff of the NIH institutes then rank the approved proposals. Priority scores are the main, but not the sole, determinants of funding: An estimated 1 to 2 percent of proposals are funded because of their particular relevance to a pressing health need, the need to start research in areas of future importance, a desire for balance in the portfolio of grants supported by an institute, ethical considerations, or importance to NIH program needs (26). Roughly one in five grant applications is referred to more than one institute by the study section (11). A small proportion of applications is funded by more than one institute; typically, however, they are funded by one institute or are not awarded.

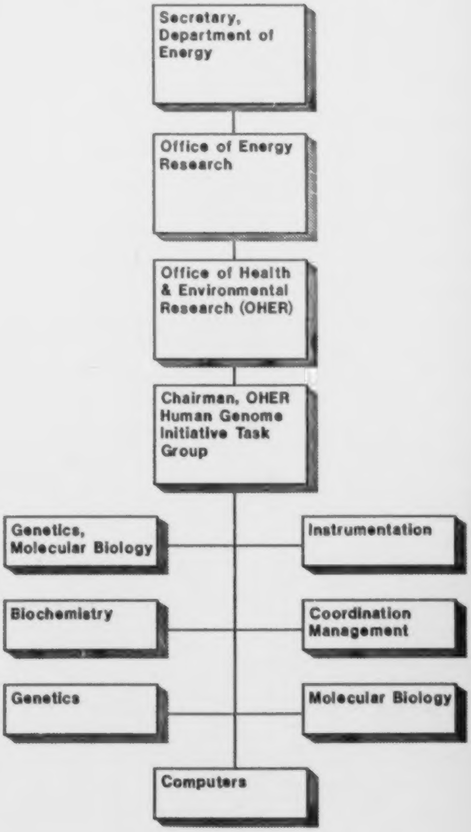
Contracts and special programs also receive peer review, usually through program review committees organized by the institutes or divisions. The intramural research programs are reviewed by non-NIH scientists who serve on boards administered by the institutes. Special review committees are also constituted by the institutes to review center or program grants.

In its 1984 report on the organization of NIH, the Institute of Medicine noted that "the genius of the institution in shaping scientific excellence to health needs is found in the interplay between the categorical research institutes and the disciplinary study sections" (11). This statement refers to the fact that except for NIGMS, the NIH institutes focus on a category of diseases or organ systems. Study sections, in contrast, are composed of scientists from a particular discipline or area of expertise (e.g., genetics, pharmacology, pathology). These may overlap, but they often do not. Institutes consider applications from different study sections, sometimes from as many as 18 (11). In 1986, there were 2,700 scientists and lay representatives serving on 155 review committees at NIH (4,26).

DEPARTMENT OF ENERGY

Much of the attention devoted to mapping and sequencing the human genome can be traced to activities in the Department of Energy. DOE has already begun a program of targeted research on the human genome—the Human Genome Initiative—to construct physical maps of several human chromosomes and to develop relevant technologies. The part of DOE responsible for the Human Genome Initiative is the Office of Health and Environmental Research in the Office of Energy Research (see figure 5-2).

Figure 5-2.—Organization of DOE


Secretary, Department of Energy
Office of Energy Research
Office of Health & Environmental Research (OHER)
Chairman, OHER Human Genome Initiative Task Group
Genetics, Molecular Biology
Instrumentation
Biochemistry
Coordination Management
Genetics
Molecular Biology
Computers

The Office of Health and Environmental Research

The history of the Office of Health and Environmental Research (OHER) goes back to the Manhattan Project of World War II, which was organized to develop fission bombs. OHER began as the Health Division, started in 1942 by Nobel laureate Arthur Holly Compton, a physicist at the University of Chicago. The division focused on protecting people from the effects of radiation and on the use of radioactive chemicals in medicine and biomedical research. The research base was broadened to include fossil fuels and renewable energy sources by the Energy Reorganization Act of 1974. These functions were retained when the Energy Research and Development Administration became the Department of Energy in 1977 and OHER was established (21). The primary mission of OHER has been to study sources of radiation, pollution, and other environmental toxins (particularly those related to the generation of energy), to trace them through the environment, and to determine their effects. Another mission is to exploit the resources of DOE-administered national laboratories to the maximum benefit of the nation.

Research at OHER is conducted largely through the system of national laboratories. There are eight general-purpose laboratories that conduct OHER research as well as research in physical sciences and mathematics, and there are nine dedicated OHER laboratories located near national laboratories or universities. In addition, OHER supports research at 100 universities and research centers.

OHER's involvement in the human genome debate is traced by its former director, Charles DeLisi, to an idea that occurred to him late in 1985 when he was reading a draft of the OTA report *Technologies for Detecting Heritable Mutations in Human Beings* (23,27). He realized the importance of having a reference human sequence for OHER's work. Subsequent discussions disclosed that researchers at the Lawrence Livermore and Los Alamos National Laboratories were thinking about ordering DNA clones to make a physical map as an extension of ongoing work. Robert Sinsheimer, chancellor of the University of California at Santa Cruz, had hosted a workshop on the feasibility of sequencing the human genome the previous year and was very interested. During this period, Nobel laureate Renatto Dulbecco published a brief article in *Science* urging that the human genome be sequenced (8).

DOE sponsored the Human Sequencing Workshop in Santa Fe, New Mexico, in March 1986, and DeLisi outlined a three-point strategy in a May 1986 memo: 1) to produce a set of overlapping DNA clones related to a physical map of human chromosomes (see ch. 2), 2) to develop high-speed automated sequencing methods, and 3) to improve methods for computer analysis of map and sequence information. DOE also funded a second workshop, Exploring the Role of Robotics and Automation in the Decoding of the Human Genome, in January 1987. Funding for the DOE initiative, based on the three-pronged attack, began in fiscal year 1987, with $4.2 million going to 10 projects at three national laboratories and at Harvard and Columbia Universities (6). DOE plans to spend $12 million on human genome projects in fiscal year 1988 and has requested $18.5 million for 1989. A special appropriation of $12.7 million was added to the Office of Energy Research budget to construct a building for an Institute of Human Genomic Studies at Mount Sinai Medical Center in New York. This resulted from a congressional initiative, and operations of that institute are not part of human genome projects sponsored by DOE. The reason for the name of the institute is unclear, but the institute will apparently house clinical genetic services for its region (24).

OHER projects include assembling an ordered set of overlapping DNA clones spanning human chromosome 19 at the Lawrence Livermore National Laboratory, making a similar clone set for chromosome 16 at the Los Alamos National Laboratory (using somewhat different techniques), and constructing a physical map of chromosome 21 and chromosome X at Columbia University. Efforts in 1988 will expand to include more university groups and the Lawrence Berkeley National Laboratory and other national laboratories. An effort to sequence the genome of the bacterium *Escherichia coli* by a new method is being supported at Harvard University. Other projects include construction of DNA clones covering the full set of human chromosomes (not cataloged in order) and development of new technologies for sequencing, detecting, and analyzing DNA.

Early enthusiasm for mapping and sequencing at the national laboratories stemmed largely from existing OHER projects. In one set of projects, laser-activated cell sorting was used to separate individual chromosomes. Cell sorting began naturally in the national laboratories because of easy access to high-technology instrumentation and a multidisciplinary blend of scientists, including biologists, chemists, physicists, computer scientists, engineers, and mathematicians. The first fluorescence-activated cell sorter was developed at the Los Alamos National Laboratory. This instrument was used at the Lawrence Livermore and Los Alamos National Laboratories to sort human chromosomes, and these chromosomes were used to produce sets of DNA clones. The effort was divided into two phases.

The first phase was to make sets of small fragments of cloned DNA (up to several thousand base pairs) in lambda phage. This phase has been completed, and the clone sets have been turned over to the American Type Culture Collection in Rockville, Maryland. (Preparation of the clones is funded by DOE; storage and distribution of the clone sets are funded jointly by NICHD and DRR. DRR also supports the cell-sorter facility at the Los Alamos National Laboratory.) The second phase is to develop clone sets of up to 45,000 base pairs using cosmids and other vectors (see ch. 2 for details). The next logical step is to order the clone sets.

In future years, DOE plans to expand its efforts substantially. A report recently written by a subcommittee of the Health and Environmental Research Advisory Committee is the main public planning document for DOE work on human genome projects.

Health and Environmental Research Advisory Committee Report

The Health and Environmental Research Advisory Committee (HERAC) is a group of scientists from universities, national laboratories, and private corporations which reports to the Director of the Office of Energy Research. Its main function is to advise the Director of OHER on the scientific program supported by OHER. In late 1986, HERAC formed a subcommittee on the human genome to make recommendations about DOE's Human Genome Initiative. The subcommittee was chaired by Ignacio Tinoco and included members from the Howard Hughes Medical Institute, universities, biotechnology companies, and one scientist from a national laboratory. The subcommittee's document was approved and was submitted by HERAC to Alvin Trivelpiece, then Director of the Office of Energy Research, in April 1987 (25).

The subcommittee report urges DOE to develop two important tools for research in molecular biology: a reference human DNA sequence and the means to interpret and use it. These would be created by a new research program divided into two stages. The first phase (5 to 7 years) would focus on:

- assembling ordered DNA clone sets of the human chromosomes;
- locating genes and other markers on a physical map based on these sets;
- producing sequences of selected clones and distributing that information;
- developing new techniques for mapping and sequencing;
- applying automation and robotics to mapping and sequencing;
- creating computational and other methods for identifying genes;
- finding new algorithms for analyzing DNA sequences; and
- establishing computer facilities, databases, materials repositories, networks, and other resources to promote use of the methods and resources produced by the projects.

Budget recommendations for the first phase are noted in table 5-3. The second phase would provide a complete sequence for each human chromosome and would make new technologies available for use in addressing the central questions of medicine and biology.

The subcommittee recommends that the work be widely distributed among national laboratories, universities, and companies because of the "highly creative nature" of the science needed to meet the objectives. The research program would include work by many small groups funded through investigator-initiated grants, as well as larger multidisciplinary centers or consortia. The report also recommends that DOE establish a two-tiered system of peer review: one or two initial review committees to assess technical merit and feasibility, and a policy committee to determine overall strategy, develop policy, and oversee scientific review.

Table 5-3.—Budget Proposed for DOE Human Genome Initiative (millions of dollars)

Fiscal year	Amount that year	Cumulative amount
1988	20	20
1989	40	40
1990	80	140
1991	120	260
1992	160	420
1993	200	620
1994	200	820
1995	200	1,020

SOURCE: Subcommittee on the Human Genome, Health and Environmental Research Advisory Committee, *Report on the Human Genome Initiative*, prepared for the Office of Health and Environmental Research, Office of Energy Research (Germantown, MD: Department of Energy, April 1987).

The subcommittee urges DOE to ensure that the results of the projects be in the public domain and that the efforts be made in cooperation with those of other agencies in the United States and abroad, within the constraints of Federal law governing technology transfer and concern for national competitiveness in biotechnology.

A broad-based research program to foster development of technology is outlined, followed by the rationale for DOE involvement, namely: 1) the historical relation to ongoing work at the national laboratories; 2) DOE's experience with directed research programs (as opposed to the much larger and more diverse human and animal research supported by NIH); 3) the relation to the mission of OHER (in assessing mutational damage from radiation and environmental exposure or developing new energy resources); and 4) access to multidisciplinary teams in the national laboratory system. The potential utility of DNA sequencing for monitoring exposure to radiation and toxic chemicals is noted as a principal reason for developing sequencing technologies.

The primary justification for the new initiative is its potential utility. The technologies and information deriving from it would make future research more efficient (less costly and more powerful), would directly improve human health, and would aid economic growth of industries dependent on biotechnology. A final section of the report warns that, although the program is of the highest priority, it should not be permitted to hinder worthwhile ongoing programs, including research on nonhuman organisms. Concern that a large new program at DOE would impede development in other fields is countered with the observation that large new sums of money have already been introduced into molecular biology: HHMI has increased its annual spending on biomedical research by over $150 million during the last decade, with primarily beneficial results. The subcommittee ends by stating its opposition to creating any large, inflexible organization to execute or supervise the work.

NATIONAL SCIENCE FOUNDATION

The blueprint for the National Science Foundation (NSF) grew from the report *Science—The Endless Frontier*, written by Vannevar Bush in 1945 (2). The original ideas for NSF, as propounded by Bush and Senator Harley Kilgore, were modified by postwar events and eventually led to legislation creating the foundation in 1950. The principal purpose of the NSF was to continue the Federal Government's role in sponsoring basic research, a role that developed during World War II (9,14). Biology at NSF is supported through its Directorate of Biological, Behavioral, and Social Sciences. In fiscal year 1987, NSF spent an estimated $32.7 million on research related to gene mapping and sequencing. Of this amount, only $200,000 went for focused projects on gene mapping and sequencing of nonhuman organisms; the bulk was for basic research ($13.7 million) and for the research infrastructure, such as development of methods, new scientific instruments, databases, and repositories and support of instrumentation centers ($19 million). Planned spending for 1988 was $37.9 million. These figures are part of the $206 million spent by NSF in support of biological science in fiscal year 1987, out of the total NSF budget of $1.62 billion (12).

NSF supports primarily basic research in all sciences. Support of basic research grants is the largest single component of NSF funding related to human genome projects. In recent years, NSF has increased its emphasis on engineering and technology development (e.g., it partially supported development of the California Institute of Technology's DNA sequenator). In 1987, NSF announced a Biological Centers Program intended to stimulate the growth of knowledge in biological research areas important to the continued development of biotechnology. Support for these centers, estimated at $12 million for fiscal years 1987 and 1988, constitutes the second largest component of NSF funding of genome-related activities. A center for bioprocess engineering has been functioning at the Massachusetts Institute of Technology for several years. (NIH has also supported this center, for research training.) Two types of centers are to be created under the Biological Centers Program: One will focus upon sharing capital-intensive instrumentation and developing new instruments; the other will host large-scale multidisciplinary research. Either could be used by groups mapping and sequencing various organisms. NSF also sponsors a program on biological instrumentation and funds individual grants for basic biological science. Although the NSF budget for biology is small relative to its support for other areas and to DOE and NIH support, it nonetheless supports mapping and sequencing through bioengineering, basic biology research, and the centers programs.

NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards (NBS) was created in 1900 as the National Standardizing Bureau. It is part of the Department of Commerce, and its primary mission since its inception has been to develop standards in scientific and technical fields in order to facilitate industrial progress and to prevent incompatibilities that could hamper research or technological applications. NBS also has a program of research in methods and instrumentation that has grown naturally out of tracking diverse and rapidly advancing technologies. Its main technical expertise lies in the physical, chemical, and information sciences, but it is now developing expertise in biotechnology. It has joined with the Montgomery County Government and the University of Maryland, for example, in support of the Center for Advanced Research in Biotechnology in Gaithersburg, Maryland.

NBS has been suggested as a candidate agency for quality control and research on measurements for DNA mapping and sequencing. This would give it the function in biology that it has for physics and chemistry but would entail a considerable expansion of its expertise and resources devoted to molecular biology. Its role could include checking data for accuracy, assessing the accuracy of the machines used in the multicenter mapping and sequencing efforts, and setting standards for the reporting of results. NBS might also conceivably develop technical standards for automated machines and computers used in creating or analyzing data about DNA. If NBS undertakes a function in quality control and standard setting, it will need close collaboration with NIH and DOE, where the bulk of expertise currently lies.

CENTERS FOR DISEASE CONTROL

The Centers for Disease Control (CDC) are situated in the Public Health Service of the Department of Health and Human Services. The main offices are located in Atlanta, Georgia. CDC is the Nation's primary resource for tracking the incidence and prevalence of diseases and for intervening to thwart the spread of infectious agents and preventable diseases. Related to this mission, CDC maintains databases, disseminates information, and provides materials widely used in clinical research.

CDC could have a role in quality control and in monitoring scientific activities involved in mapping and sequencing the human genome. It has performed this function in the past, through its Lipid Standardization Program. This program began more than 25 years ago to provide quantitative measurements for laboratories engaged in lipid research related to diseases of the heart and blood vessels. Since the program was initiated, over 500 national and international laboratories have received and analyzed reference materials provided by CDC (3). Quality control and standard setting may become important as map and sequence data become more plentiful and as more laboratories come to rely on a common set of data. If such measures prove necessary, CDC is a possible agency for determining or confirming the chromosomal location or origin of DNA fragments or for orienting new DNA fragments on the emerging physical maps. If this function were to be undertaken by CDC, close communication with NIH and DOE would be necessary.

DEPARTMENT OF DEFENSE

While biological research is not the main mission of the Department of Defense (DoD), some components of mapping and sequencing DNA might be shared with or conducted by various components of DoD. Each military service (particularly the Army and Navy) conducts some research in biology, primarily that related to the health needs of military personnel or to defenses against chemical and biological warfare. DoD reports that all such research is unclassified: Much of it is conducted at military facilities or contractor-administered laboratories, but some of it is conducted in universities as well. DoD supports some generally useful resources in biomedical research.

The Armed Forces Institute of Pathology (AFIP) is an international treasure house of tissue samples and microscopic slides spanning the full range of human disease. Its tissue collection is used by pathologists and biomedical researchers throughout the world. AFIP began as the Army Medical Museum in 1862. It became the AFIP in 1949, when the Navy and Air Force joined with the Army in support of it, and the role of the institute has expanded steadily since then. Today AFIP constitutes the largest organization of research and diagnostic pathologists in the world. The institute has received more than 2.2 million cases (tissue or slides from patients) from over 50,000 pathologists affiliated with more than 19,000 hospitals and clinical facilities. AFIP's unique capabilities as a tissue repository have been expanded to include modern storage techniques. Through further expansion of its capabilities, the staff and facilities of AFIP could be used as a national tissue repository and assessment center for the full spectrum of human diseases. The institute could play a role in linking map and sequence data to human diseases. The availability of systematically classified human tissues could facilitate development and testing of medical products and diagnostic methods to probe the molecular basis of various diseases.

The military biomedical research community would have an interest in map and sequence data because investigations of the effects of chemical and biological weapons would include the study of genes that are particularly vulnerable to attack and the construction of vaccines or other defensive measures.

OFFICE OF SCIENCE AND TECHNOLOGY POLICY

The Office of Science and Technology Policy (OSTP) is headed by the President's Science Advisor. OSTP's primary responsibility is to advise the President on science policy, on matters where scientific or technical information is relevant to Federal policy decisions, and on national policies for technology development. OSTP can on occasion form coordinating councils under the Federal Coordinating Council for Science, Engineering and Technology. An example of OSTP coordination in life sciences is the Biotechnology Science Coordination Committee, which started as an OSTP initiative responsible primarily for devising guidelines for regulation of biotechnology products. Representatives of OSTP followed the human genome debate and spoke at several national meetings in 1986 and 1987.

OSTP recently announced plans to reorganize its oversight of life sciences. It plans to form a Committee on Life Sciences for interagency communication and coordination, and it tentatively plans to establish subcommittees on specific topics. Genome projects have been noted as likely to necessitate such a subcommittee, although the exact role and composition of it is not yet determined (7).

DOMESTIC POLICY COUNCIL

The Domestic Policy Council (DPC) is a cabinet-level group that reviews government activities. David Kingsbury of NSF, as acting chairman of the Biotechnology Working Group, gave a brief presentation on human gene mapping to the DPC in February 1987. An interagency subcommittee of this working group, chaired by NIH Director James Wyngaarden, was formed and met in May 1987 to exchange information on agency activities. NIH, DOE, NSF, the Food and Drug Administration, the U.S. Department of Agriculture, the Environmental Protection Agency, and the Office of Management and Budget were represented. The purpose of the subcommittee was to minimize duplication of effort among the agencies and to promote interagency communication. The subcommittee has subsequently been disbanded, to be replaced by the OSTP group noted above. The DPC will continue to keep abreast of developments on genome projects through the President's Science Advisor, who will administer the OSTP group (1).

OFFICE OF MANAGEMENT AND BUDGET

The Office of Management and Budget (OMB) monitors and coordinates the annual budget process for executive agencies of the Federal Government and oversees management of the agencies. Each year, every Federal agency prepares a budget request that is reviewed within the agency and then submitted to OMB. OMB reviews the requests and develops a budget for the President; this budget is submitted to Congress in January for the fiscal year beginning that October (although the process is late for fiscal year 1989 because of delay in passing the 1988 budget). OMB's budget-coordinating function places it in the position of arbiter among different agencies if there are conflicting priorities or potential duplications. By this mechanism, and by monitoring other activities in multiple departments, OMB can encourage communication and coordination of activities. OMB has one budget officer for NSF, another for NIH, and a third for DOE. These officers are responsible for other agencies as well, and the activities related to human mapping and sequencing constitute only a small fraction of their total budget responsibility. Two officers in the OMB science office have taken primary responsibility for tracking the human genome budget submissions of all agencies (20).

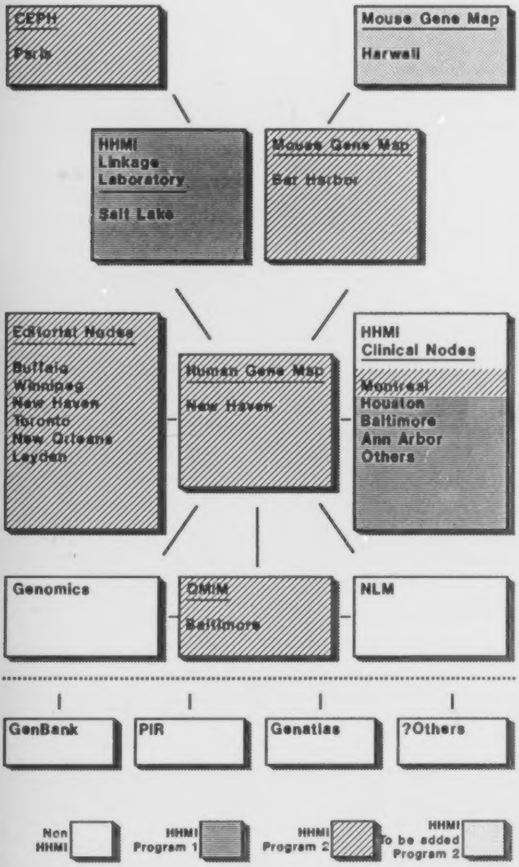
HOWARD HUGHES MEDICAL INSTITUTE

The Howard Hughes Medical Institute (HHMI) was created in 1953 by aviator-industrialist Howard Hughes. It is a medical research organization with an endowment of approximately $5 billion. HHMI has increased its research funding dramatically over the last decade, from roughly $15 million in 1977 to approximately $240 million in 1987.

HHMI operates three programs (see figure 5-3). The first and largest is scientific research in 27 laboratories located in hospitals, academic medical centers, and universities throughout the United States. The second program, which supports the first research program and is integrated with it, includes a genome resources project, a research training program for medical students (jointly with NIH), and sponsorship of HHMI meetings and reviews. A third program will provide $500 million over the next decade through grants and special programs to support education in the medical and biological sciences.

Under its first program, HHMI conducts research in five basic scientific areas: genetics, im-

Figure 5-3.—Howard Hughes Medical Institute Genomics Resources


CEPH
Paris
Mouse Gene Map
Harwell
HHMI
Linkage
Laboratory
Salt Lake
Mouse Gene Map
Bar Harbor
Editorial Nodes
Buffalo
Winnipeg
New Haven
Toronto
New Orleans
Leyden
Human Gene Map
New Haven
HHMI
Clinical Nodes
Montreal
Houston
Baltimore
Ann Arbor
Others
Genomics
OMIM
Baltimore
NLM
GenBank
PIR
Genatlas
?Others
Non HHMI
HHMI Program 1
HHMI Program 2
HHMI o be added Program 2

munology, neuroscience, cell biology and regulation, and structural biology. Several HHMI investigators are involved in genetic mapping and related computational research, physical mapping, and medical genetics. The principal HHMI centers for genetics are located at the University of Utah (Salt Lake City, Utah), the Baylor College of Medicine (Houston, Texas), the University of Michigan (Ann Arbor, Michigan), the University of California (San Francisco, California), The Johns Hopkins University (Baltimore, Maryland), the University of Pennsylvania (Philadelphia, Pennsylvania), Brigham Hospital (Boston, Massachusetts), and Childrens' Hospital (Boston, Massachusetts). HHMI estimates that it expended $40 million for genetics research in 1987, of which $2 million to $4 million were devoted to finding and using DNA markers and constructing genetic maps.

The HHMI genome resources project has a current annual budget of over $2 million. Through that project, HHMI supports nonsequence databases relating to human genetics, including the Human Gene Mapping Library (New Haven, Connecticut) and the *On-Line Mendelian Inheritance in Man* (Baltimore, Maryland) (see figure 5-2). HHMI also helps maintain a mouse genetics database at Jackson Laboratory (Bar Harbor, Maine) and collaborates with the Center for the Study of Human Polymorphism (CEPH) headquartered in Paris, France. CEPH is a critical collaborative institution that links several large groups working on construction of human genetic maps. HHMI has participated in a large number of meetings on the human genome, including one it sponsored directly in July 1986 at NIH; at that workshop, strategies and policies for mapping and sequencing were discussed. HHMI partially supported the ninth Human Gene Mapping Workshop, held in Paris in September 1987, which compiled data on international gene mapping activities since the previous meeting in 1985.

NATIONAL RESEARCH COUNCIL

The National Academy of Sciences was established by Congress in 1863. President Lincoln signed the law that brought it into existence. The National Research Council (NRC) was established in 1916 to provide advice to the Federal Government about issues involving science. The principal impetus was the increasing relevance of science to preparations for World War I. The National Research Council now conducts many studies on issues relating to science and technology. It is organized into several disciplinary groups.

In August 1986, a group of scientists interested in issues surrounding genome projects met in Woods Hole, Massachusetts. This group agreed to formulate a proposal for a study by the NRC, which was presented to and approved by the Basic Biology Board of the Commission on Life Sciences in September 1986. A committee of distinguished scientists, chaired by Bruce Alberts, was appointed to consider the scientific issues connected with genome projects. The committee on mapping and sequencing the human genome held several public meetings in 1987 and released its report in February 1988 (19).

The Alberts committee was composed of scientists of different backgrounds, with varying degrees of direct involvement in mapping and sequencing projects and with initially divergent views on genome projects. The committee reached consensus on several points during the course of its deliberations. The committee concluded that mapping, sequencing, and understanding the human genome merited a special effort funded and organized specifically for this purpose.

The committee's report recommends that the projects should begin with "a diversified, sustained effort to improve our ability to analyze complex DNA molecules," with a "focused effort that emphasizes pilot projects and technological development." It lists the specific types of maps that would be useful as early genome projects, notes the importance of mapping and sequencing genomes of nonhuman organisms, and stresses the need for thorough peer review. The proposed projects differ from ongoing research by focusing on methods that would improve mapping, sequencing, analyzing, or interpreting the biological significance of information in the human genome by five- to ten-fold. The committee also notes the need for central databases, repositories, and quality control facilities.

Research projects that merit special support are explained in some detail. The committee favors development and refinement of techniques in the early years, with most support going to work on mapping large genomes. One specific goal in early years, for example, would be to enable sequencing of 1 million continuous base pairs. The need for technological progress is noted for several additional areas: to isolate chromosomes, to create cell lines, to clone substantial portions of DNA from genomes of whole organisms, to clone DNA in large fragments, to isolate large DNA fragments, to order DNA clones derived from genomes, to automate many steps involved in mapping and sequencing DNA, and to improve the collection, storage, dissemination, and analysis of information and materials. Administration of these centralized functions would be conducted by a scientific advisory board, including at least one full-time scientist appointed as chairman. This scientific advisory committee would also serve to advise the agencies and to act as the focal point for international cooperation.

The committee recommended that $200 million per year be appropriated specifically for genome projects, increasing to this level over the first 3 years. In the first 5 years, this might be spent to fund work at 10 medium-sized multidisciplinary centers and to support a program of grants to many more small research groups. An estimated 1,200 scientists would be involved, with roughly half located at the multidisciplinary centers. The research component would account for $120 million per year. The remainder of the budget would be used for construction ($55 million per year initially, decreasing in later years) and to pay for the repository, database, quality control, and administrative functions of the scientific advisory committee ($25 million per year). The funding for construction in early years would be reassigned to production of maps and sequence data as technologies matured.

A majority of the committee recommended that a single agency be designated and given funding to lead the effort. Other options were also discussed, including an interagency structure much like the task force option discussed in the next chapter. A final option was to have an interagency body for planning and funding, but a single agency for administration.

PRIVATE CORPORATIONS

Private corporations in several fields have expertise relevant to mapping and sequencing the human genome. Many instruments first developed in academic or national laboratories are now produced commercially. Pharmaceutical and biotechnology companies could use map and sequence information to develop new products, and many are themselves developing research techniques. Companies that market scientific instruments are also keenly interested. The role of the private sector appears to be primarily to:

- advise in planning the mapping and sequencing research program,
- commercialize products that result from the research, and
- ensure that technology transfer from federally funded research projects to commercially exploitable products is smooth and rapid.

Private corporations view human genome projects, with a few exceptions, as long-term research that is best supported by the Federal Government. Corporations are unlikely to lend financial support to a national program to map and sequence the human genome, although they might well invest in particular projects that involve development of technology. Private firms could perform specified functions under contract from the Federal Government (e.g., genetic mapping, physical mapping, DNA preparation, or DNA sequencing) once the technologies are available.

Several American companies already produce DNA sequenators and other analytical instruments used in mapping and sequencing projects. The Lawrence Livermore National Laboratory group, which is constructing ordered sets of DNA clones under DOE sponsorship, is modifying an instrument initially designed for DNA sequencing by Applied Biosystems of Foster City, California. Private corporations have likewise participated in building the existing genetic map of human chromosomes. Collaborative Research and Integrated Genetics are two companies based in the Boston area that have contributed substantially to the effort to find new DNA markers and to link those markers to human diseases. A few biotechnology companies have been at the forefront in developing automated technologies for handling DNA. The Genetics Institute (Cambridge, Massachusetts), for example, developed a robotic system that extracts DNA from bacteria and cells.

At least two companies—the Genome Corp. in Boston and SeQ, Ltd., in Cohasset, Massachusetts—are being started specifically to map and sequence the human genome. These companies plan to construct a physical map and subsequently sequence the human genome over the next decade, using private funds. They would offer access to the materials and to the map and sequence information for a price. The process would be much like that used currently by researchers, who pay repositories for DNA clones, probes, and vectors or who pay companies for enzymes and other materials used in molecular biology. The argument behind this is that, while each laboratory could conceivably develop the information independently, it is cheaper and faster simply to buy it from a private firm that has developed it already. Those purchasing the information would be free to use it, but not to copy or sell it.

Private corporations could also play a role in the development of technology related to mapping and sequencing. This could include company access to government facilities, exchange of corporate and academic personnel, multicompany consortia, individual corporate agreements with universities or national laboratories, or some combination of these.

Members of the Industrial Biotechnology Association were recently polled regarding their support of Federal initiatives in mapping and sequencing the human genome. Those responding indicated that:

- The work should be funded entirely by the Federal Government and should not interfere with ongoing biomedical research.
- NIH, DOE, and NSF should all participate (NIH should take the lead).
- A national planning committee, composed of 50 percent university scientists, 30 percent government representatives, and 20 percent industry representatives, should be set up.
- Work should be carried out at dispersed university and federally supported laboratories, not a center created for the purpose.
- International cooperation should be encouraged if it does not entail delays.
- Physical mapping should precede sequencing.

Respondents clearly support a role for industry in planning and using the results of mapping and sequencing projects, while indicating that the Federal Government should pay the bill.

PRIVATE FOUNDATIONS

Several private foundations support research in human genetics. These include such disease-oriented foundations as the March of Dimes, the Hereditary Disease Foundation, the Muscular Dystrophy Association, and the Cystic Fibrosis Foundation. Other foundations support work on human genetics as part of a broader research program, among them the American Cancer Society, the American Heart Association, and the Alzheimer's Disease and Related Disorders Association. These foundations, while relatively small in total funding, often act as catalysts in focusing research on a problem of particular interest. They are also highly effective at publicizing research results, educating the public about the consequences of disease, and generating public support for biomedical research.

SUMMARY

NIH, DOE, HHMI, and NSF have already made substantial commitments to projects related to the study of the human genome. Government activities are currently being coordinated by informal communication among the agencies. A previous coordinating group under the Domestic Policy Council will likely be replaced by one organized under the Office of Science and Technology Policy in the White House, with budget submissions coordinated by the Office of Management and Budget. Each research agency has its own means of funding research and providing peer review of programs. NIH and DOE have created special planning groups to review genome projects. NIH funding is over $313 million each year for human and nonhuman research involving mapping or sequencing. In 1987, NIH announced two new programs in methods development; it has budgeted $17.2 million for those projects in 1988 and requested $28 million for 1989.

In 1987, DOE allocated $4.2 million for 10 projects on physical mapping and technology development. It plans another $12 million in 1988 and has received recommendations from an outside scientific panel to ask for over $1 billion over the following 7 years. The former director of OHER stated that at least half the funds would be distributed to researchers at universities and research centers other than national laboratories (3) and that the work will be reviewed prospectively and retrospectively by peers. DOE officials have stated that their budget requests will be more modest than those recommended.

NSF spent over $32.7 million on research related to genome projects in 1987, although only $200,000 was considered to be for genome projects per se. NSF's new Biological Centers Program is likely to be relevant to genome projects, particularly those involving new instrumentation. HHMI funded $40 million of genetics research in 1987, including several million for construction of genetic maps. HHMI also administers and funds a genomics resource program of $2 million annually to support databases and other elements of the research infrastructure.

To date, actions of the principal organizations can be described as cooperative. NIH and DOE have supported many joint efforts related to human genome projects. The GenBank® database has been administered by NIH and located at a DOE-supported national laboratory for over 5 years, and the two agencies also jointly support DNA clone and probe repositories, computer analysis methods, and flow-sorting facilities. NIH and DOE sponsored a meeting on database and repository needs of human genome projects in August 1987, and there has been an exchange of project officers and extensive informal cooperation among staff at NIH, DOE, NSF, and other executive agencies.

The strengths of NIH and DOE are more complementary than competitive. Each believes it could successfully mount and sustain the scientific and technical effort necessary for the contemplated mapping and sequencing projects. Both support relevant work already, although with different emphases. A decision to delegate the entire effort to one agency would require that current efforts in other agencies be shut down.

The mapping and sequencing effort will more likely continue to include NIH, DOE, NSF, HHMI, and other agencies and organizations covered in this chapter. The key question then becomes how much and which part each agency should perform. Such decisions will be made in a general sense by Congress, through authorization and appropriation, with more detailed planning left to executive agencies. There may be informal cooperation or some more formal means of coordinating the planning and execution of agency projects under OSTP. Joint nongovernment advisory groups could be formed to bring in expertise from academia and industry. There are many options for organizing an interagency effort and for incorporating outside advice into research planning. These issues of organization and advisory structure are discussed in chapter 6.

CHAPTER 5 REFERENCES

1. Bledsoe, R., White House Domestic Policy Council, personal communication, December 1987.
2. Bush, V., *Science—The Endless Frontier: A Report to the President* (Washington, DC: reprinted by the National Science Foundation, 1980).
3. Chen, A., Centers for Disease Control, personal communication, August 1987.
4. Committee Management Staff, National Institutes of Health, *NIH Public Advisory Groups,* NIH Pub. 87-10 (Bethesda, MD: NIH, October 1986).
5. DeLisi, C., comments at a meeting of the Human Genome Subcommittee, Health and Environmental Research Advisory Committee, Department of Energy, Denver, CO, December 1986.
6. Department of Energy, material submitted to the Domestic Policy Council, May 1987.
7. Dorigan, J., White House Office of Science and Technology Policy, personal communication, December 1987.
8. Dulbecco, R., "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science* 231:1055-1056, 1986.
9. England, J.M., *A Patron for Pure Science: The National Science Foundation's Formative Years, 1945-1957* (Washington, DC: NSF, 1982).
10. Harden, V.A., *Inventing the NIH: Federal Biomedical Research Policy, 1887-1937* (Baltimore, MD: The Johns Hopkins University Press, 1986).
11. Institute of Medicine, *Responding to Health Needs and Scientific Opportunity: The Organizational Structure of the National Institutes of Health* (Washington, DC: National Academy Press, October 1984), p. 76.
12. Kingsbury, D., National Science Foundation, personal communication, October 1987.
13. Levinson, R., Directors Office, National Institutes of Health, personal communication, January 1988.
14. Mazuzan, G. F., "National Science Foundation: A Brief History," unpublished manuscript, September 1987.
15. McKusick, V.A., *Mendelian Inheritance in Man*, 7th ed. (Baltimore, MD: The Johns Hopkins University Press, 1986).
16. Miles, W.D., *A History of the National Library of Medicine: The Nation's Treasury of Medical Knowledge,* NIH Pub. 82-1904 (Washington, DC: U.S. Government Printing Office, 1982).
17. National Institutes of Health, *NIH Guide to Grants and Contracts*, vol. 16, May 29, 1987, pp. 9-14; Oct. 16, 1987, pp. 6-10.
18. National Institutes of Health, Office of the Director, background information and information provided to the Domestic Policy Council, May 1987.
19. National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
20. Noonan, N., White House Office of Management and Budget, personal communication, October 1987.
21. Office of Energy Research, Department of Energy, *Health and Environmental Research: Summary of Accomplishments*, vol. 1, DOE Pub. ER-0194, April 1984; vol. 2, DOE Pub. ER-0275, August 1986 (Washington, DC: DOE, 1984, 1986).
22. Office of Program Planning and Evaluation, National Institutes of Health, *The Human Genome*, proceedings of the 54th meeting of the Advisory Committee to the Director, Oct. 16-17, 1986 (Bethesda, MD: NIH, 1987).
23. Science and the Citizen, "Geneshot," *Scientific American*, May 1987, pp. 58-58a.
24. Smith, D., Office of Health and Environmental Research, Department of Energy, personal communication, January 1988.
25. Subcommittee on the Human Genome, Health and Environmental Research Advisory Committee, *Re-*

port on the Human Genome Initiative, prepared for the Office of Health and Environmental Research, Office of Energy Research, Department of Energy (Germantown, MD: DOE, April 1987).

26. U.S. Congress, General Accounting Office, *University Funding: Information on the Role of Peer Review at NSF and NIH*, GAO Pub. RCED-87-87FS (Washington, DC: U.S. Government Printing Office, March 1987).
27. U.S. Congress, Office of Technology Assessment, *Technologies for Detecting Heritable Mutations in Human Beings*, OTA-H-298 (Washington, DC: U.S. Government Printing Office, September 1986).

Chapter 6

Organization of Projects

CONTENTS

Page
Administrative Structures 115
  Single-Agency Leadership 116
  Interagency Agreement and Consultation 118
  Interagency Task Force 119
  Consortium 121
  Discussion 123
Advisory Structure 123
  Responsibilities 123
  Composition 124
  Structure and Funding 124
Big Science v. Small-Group Science 125
  Displacement of Higher-Priority Science 125
  Style and Efficiency 125
  Politicization 127
Summary 128
Chapter 6 References 128

Boxes

Box *Page*
6-A. Acid Precipitation Task Force 120
6-B. Midwest Plant Biotechnology Consortium 122
6-C. Quotes on Genome Controversies 126

Figures

Figure *Page*
6-1. Lead Agency 116
6-2. Interagency Agreement and Consultation 118
6-3. Interagency Task Force 119
6-4. Consortium 121

Chapter 6

Organization of Projects

"Organization is a means to an end rather than an end in itself. Such structure is a prerequisite to organizational health; but it is not health itself. The test of a health business is not the beauty, clarity, or perfection of its organization structure. **It is the performance of people**."

Peter Drucker,
Management: Tasks, Responsibilities, Practices
(New York: Harper & Row, 1974), p. 602.

ADMINISTRATIVE STRUCTURES

Chapter 5 presented the history, current involvement, and future plans of the many government and nongovernment parties interested in genome research. This chapter assumes the continued interest and participation of the current actors, and it discusses the options for organizing those actors at the Federal level.

A properly designed administrative or organizational structure for genome projects is important. As form so perfectly matches function in DNA, so the organizational form of the project should match its function and goals. These include rapid accumulation of knowledge about the genome, efficient storage and distribution of that information, and conversion of this knowledge into productive theories, tools, reference materials, and medicines. The political consequences of a poorly administered group of projects are not only failure to achieve potential intellectual and economic contributions, but also negative impacts on the organization and funding of other scientific investigations. A genome project blueprint cannot be drawn without taking into consideration the abutting structures as well as the internal constraints.

Three major funding agencies must be included in any consideration of organizational design: the National Institutes of Health (NIH), the Department of Energy (DOE), and the National Science Foundation (NSF). Nongovernmental bodies such as the National Academy of Sciences (NAS) and the Howard Hughes Medical Institute (HHMI) are already participating in organizational and advisory roles, and commercial firms anxious for sequencing technology and data seek input as well.

There are at least five possible administrative structures a human genome project could develop:

- *One agency*—a project performed exclusively by one of the expert agencies.
- *Single-agency leadership*—a project in which Congress would designate one agency to coordinate and oversee the research.
- *Interagency agreement and consultation*—a cooperative project among the agencies in which no additional authority structure would be created.
- *Interagency task force*—a project in which a committee with the authority to direct research planning among the agencies would be chartered.
- *Consortium*—a project in which the private sector as well as the Federal Government would plan research, with possible cofunding from the corporate partners.

The first alternative, a project organized and executed solely by one agency, may be dismissed as unnecessary and politically unworkable. A single-agency project could only result from cutting out others, and several agencies have already made substantial investments in genome research and related technologies. Further, the current genome infrastructure, including GenBank® and DNA clone repositories, is already interagency.

The other four proposals have unique strengths and weaknesses. For any of them to be successful, however, the administrative structure must at least organize communications at the scientific, interagency, and international levels. At most, it should be capable of planning a research program involving many partners and funding them accordingly. Congressional decisions on the organizational structure can be based on perceptions of the necessary patterns of authority, of quality and scope of experience in research and development, and of fiscal and economic priorities.

Single-Agency Leadership

One possible beginning for genome projects would be the designation by Congress of a lead agency to coordinate ongoing activities in various agencies (see figure 6-1). This option was the one favored by a majority of those on the National Research Council committee that issued a report on mapping and sequencing the human genome (18). The strengths of this organizational option derive from its clear designation of authority. Such leadership can be dynamic, and research would follow the theme established by the lead agency. A lead agency, and thus a lead administrator, focuses the project in all its aspects: It provides a communications link among researchers, domestic and foreign; a contact for media; and a target of criticism and politicking. Drawbacks to designating a lead agency are the possibility of incomplete commitment by the lead agency and the potential inability of the lead agency to command the resources of other agencies effectively. Choosing this option would necessarily entail choosing which agency should lead.

Among NIH, DOE, and NSF—the three funding agencies—NIH and DOE are the most appropriate candidates to lead a genome project. NSF is an unlikely leader because its mandate excludes the investigation of human health and disease, the ultimate focus of the projects. Further, a large-scale operation conducted by NSF would inevitably detract from other research of which NSF may be the sole patron. Such a loss would occur in each of the funding agencies involved, but NSF may be most sensitive because it funds much less biology than NIH, and the research it supports is more basic as a rule than that supported by DOE or NIH (30). NSF can contribute to genome projects by stimulating interest in automation and robotics (which it has done), in animal models of human disease (by gathering animal and microorganism sequence data for comparison), and in instrumentation (through its biology centers).

Figure 6-1.—Lead Agency

Choosing between NIH and DOE is troublesome because these agencies have complementary strengths and weaknesses. The project would have a different face with different leadership.

Because of its mandate to support investigations to improve the Nation's health, NIH dominates biomedical research. The institutes spent an estimated $313 million in 1987 for projects that involved mapping or sequencing, over $90 million of which funded projects to characterize the human genome (13,15). NIH has conducted, funded, and administered genetics research expertly for years, and the institutes would seem to be a natural home for genome projects. The theme of an NIH-led project would likely be a renewed commitment to the peer review system and to small, or cottage industry, science, with some added attention to the research infrastructure.

One great strength of NIH is its decentralized administration: Quality projects uninteresting to one institute may well be funded by another. This flexibility is achieved at a cost, however. Critics have scrutinized this process and concluded that, among other faults, it cannot support a large, directed project (28). NIH leadership can have difficulty imposing the priority decisions needed for a concerted effort. A distinction is often made between the operating styles of NIH and NASA (the National Aeronautics and Space Administration), with NASA having much greater central authority and NIH exemplifying a decentralized process for setting priorities. To manage some of the proposed genome projects, mechanisms beyond NIH's standard researcher-originated format may be required. This could "require a change in NIH's philosophical outlook and its approach," according to George Cahill of HHMI (23).

NIH could conduct a directed research program. It has done so in the past for the study of particular diseases (e.g., polio and cancer) and is now doing so for AIDS. While NIH has not previously mounted a major project to develop a set of tools for biology (as mapping and sequencing projects are often characterized), it has the funding mechanisms and expertise necessary to do so. The mapping and sequencing projects have been described as a library of information awaiting translation, and NIH administers the National Library of Medicine (see ch. 5). To facilitate special genome projects, NIH has created new study sections to review grants that focus on methods; NIH could also convene a new scientific advisory body in an existing institute to direct a focused project, set aside funds for special projects in one or more institutes, and begin new centers or multidisciplinary programs analogous to existing ones. It already has a multi-institute coordinating body to develop special initiatives like those announced in May and October 1987 for analyzing complex genomes and for informatics in molecular biology. NIH is currently in the process of establishing a mechanism for obtaining outside advice.

The high capital needs in some areas, the diverse expertise (extending beyond biomedical research) needed on some research teams, and the standardized and repetitive work of mapping and sequencing may render small research groups unable or unwilling to do such work (8). Experience at the National Cancer Institute with the Special Virus Cancer Program has suggested that the standard grant mechanism is insufficient for such tasks as the production of standardized tools, the distribution of clinical materials, and the increased coordination of investigators (33). If the institutes were to assign high priority to genome projects, those projects could conflict with other major research efforts, for example research on AIDS. Some persons, among them Ruth Kirschstein, Director of the National Institute of General Medical Sciences, have questioned whether "it would be appropriate to have a specifically targeted program that would compete with all the extraordinarily important programs NIH funds" (23). The danger is that a targeted program would become an instead-of program rather than an in-addition-to program, as was the case with the Special Virus Cancer Program (33).

As a lead agency, DOE would endow the genome project with different characteristics of organization and expertise. DOE has long supported research on human mutations and DNA damage and repair through the Office of Health and Environmental Research. The mission of OHER is to understand the effects of radiation and other means of energy generation on human health and the environment. OHER views ignorance of the genome and the inability to sequence and analyze DNA rapidly as major limitations on its research. As NIH might emphasize the human disease aspects of genome research, DOE would emphasize the investigation of mutagenesis and other areas closely related to OHER's mission. Critics have characterized OHER's rationale as "clearly impractical" (17) and "forced and . . . disingenuous" (30). But because of OHER's interest in human genetic material, DOE already has established expertise in crucial technologies such as automated chromosome and cell sorting, and in the computer storage of genetic data. DOE believes that, through its national laboratory structure, it should develop methods and tools useful to the entire community of molecular biologists (27).

The strengths and weaknesses of DOE are largely complementary to those of NIH. DOE's strength is its familiarity with the administration of focused research programs. It manages many large facilities for research in physics and chemistry—such as accelerators for high-energy physics—and the scientists whom DOE funds in these areas are among the best in the world. DOE also maintains excellent computing resources. Yet DOE does not have the same stature within the community of molecular biologists that NIH does (11,16,24,30).

The national laboratories have long provided services to the community of molecular biologists that are not provided by other agencies. The national laboratories have pioneered many high-technology instruments useful in biology: zonal centrifuges, high-pressure liquid chromatography, fluorescence-activated cell sorters, and chromosome sorting. Teams at national laboratories have prepared sets of DNA clones from individual human chromosomes, and current mapping projects are logical extensions of this work. Even though the national laboratories are not renowned for their expertise in molecular biology, some of the technology, analytical software, and new methods that need to be developed will not be in biological disciplines—they will involve engineering, physics, and mathematics, all areas of acknowledged national laboratory expertise.

DOE enjoys the reputation of being a proficient organizer of projects among government, university, and industry researchers. As an agency, it is experienced in managing large projects and disbursing large sums of money, extramurally to universities and research centers and intramurally to the national laboratories. Critics fear that, if DOE assumes leadership of genome projects, its bias toward central management will corrupt research and stifle the more traditional, perhaps more creative, cottage industry approach.

DOE's review process for genome projects would involve prospective and retrospective peer review. The degree of scrutiny is not likely to differ substantially from that at NIH. Funding through DOE would be less likely to sap other biomedical funds, but other biological research programs at DOE could suffer. Designating DOE as the lead agency would give the organizational lead to an agency that supports only a small fraction of related research and thus only a small fraction of the user community. Having DOE administratively lead all genome projects could prove unmanageable in the long term.

The controversy over which agency should lead —DOE or NIH—may be misguided. Each agency has a role to play, and discussion should focus instead on how to encourage cooperation and to ensure that the research program of one agency does not inhibit that of the other. One observer has asserted that a major directed program at NIH alone would soon be politically incorporated into the overall NIH budget and would thereafter displace untargeted research. The corollary is that "DOE could find the leadership excellence more easily than NIH could provide the budgetary insulation" (14). Nonetheless, NIH is the logical choice for lead agency if Congress chooses to designate one—its mission is most directly affected, and the scientific community now supported by NIH is by far the largest of the intended beneficiaries of genome projects. If NIH leads, then the expertise and multidisciplinary research already supported by NSF and DOE should be explicitly taken into account in future planning. Difficulties in designating a lead agency are discussed under options for action by Congress in chapter 1.

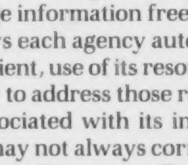
Interagency Agreement and Consultation

The lack of a lead agency implies no favored research strategy or funding mechanism, but a balanced program to take advantage of NIH, DOE, NSF, and other agencies' strengths. Agencies could be left to themselves to cooperate and communicate among themselves and with other interested organizations in the United States and abroad (see figure 6-2). A group of agency principals—agreed to by the agencies or under the Office of Science and Technology Policy—could meet to achieve these goals and exchange details of research directions and developments.

An interagency agreement and consultation framework eschews any formal creation of authority and relies on the good will of the participants to exchange information freely. Such an arrangement allows each agency autonomous, and presumably efficient, use of its resources and permits each agency to address those research topics most closely associated with its institutional interest. Interest may not always correspond to expertise, however, and it may conflict with or overlap other agencies' programs. This would act against one ostensible goal of the cooperative effort—to streamline projects by eliminating unnecessary duplication of research. An informal or ad hoc framework may also be inappropriate for very expensive, long-term projects because evolving and potentially diverging priorities may diminish rapport among the agencies.

A communications and consultation committee could be responsible for these cooperative, com-

Figure 6-2.—Interagency Agreement and Consultation



munications, and streamlining functions, but the institutional focus would be scattered and the project would exist without clear leadership. Although in the best scenario such a committee would be completely abreast of all the domestic research, it might be too diffuse a body to support international organization of a project.

Decentralized authority is not without benefits, however, for pluralism of funding sources and flexible, decentralized organization are strengths of American science. Genome projects may be compelling enough to turn the organizational gears without creating a special bureaucracy for the task. A cooperative effort also permits a flexible mix of funding options, and each agency would retain control over its research planning.

The subcommittee on the human genome of the Biotechnology Working Group of the Domestic Policy Council acted as "a mechanism for exchanging information . . . [with] the right people at the right level," according to David Kingsbury of NSF (23). This coordinating group will now be located under a life sciences committee at the Office of Science and Technology Policy. Such a group of government administrators facilitates interagency communication but may not address other needs. A purely government body is open to the criticism that scientists and not government administrators must provide direction or at least participate directly in planning (31). This conflict is similar to that found in creating an advisory body, which generally reflects the question of how much influence scientists should have on the science policy process (discussed below).

The merit of informal agreement and consultation is that each agency would have the flexibility to follow its own research agenda. Agreement and consultation would not require legislation by Congress and would make interagency cooperation a matter of congressional oversight. A disadvantage is that flexibility may be achieved at the cost of clear authority and accountability. Further, there might be no mechanism for resolving conflicts among agencies. The appropriateness of informal interagency cooperation turns on a judgment of which is more efficient—a directed and planned effort or a pluralistic and decentralized process.

Interagency Task Force

A genome initiative might require more active leadership than that described above. An interagency task force dedicated to pursuing the genome project and wielding some authority over funding and research might provide such leadership (see figure 6-3).

The task force could be constituted much like an interagency committee, with principals from the participating agencies; however, the task force would possess authority in certain areas, such as gathering of information from participating agencies, preparation of reports, formulation of recommendations, and interagency planning. It could design and direct a genome project, drawing on each of the participating organizations (see box 6-A).

A task force would be much like a lead agency in its ability to draw the attention of foreign researchers, the media, and domestic political interests. And like a lead agency, the task force would present a central character—its chairperson —who would act as spokesperson for the project. If the chairperson of the task force were selected from the agency representatives, however, the appointment would likely carry with it the same kind of political difficulties as selecting a lead agency.

Establishing a functional authority may require substantial political investment, but the cost of subsequent decisions is negligible because they can be immediate and final. With a committee authorized only to facilitate communication, a dilemma in the assignment of a particular research project, for example, could be costly in any number of ways, from the time it takes to reach a cooperative solution to the money required to duplicate the research should no equitable distribution

Figure 6-3.—Interagency Task Force

Box 6-A.—Acid Precipitation Task Force

The Acid Precipitation Act of 1980 (Public Law 96-294), Title VII of the Energy Security Act, established a 10-year program to reduce or eliminate the sources of acid precipitation. To implement this program, Congress mandated the formation of the Acid Precipitation Task Force, composed of members from the national energy laboratories, the agencies, and four presidential appointees, and chaired jointly by representatives from the National Oceanic and Atmospheric Administration (NOAA), the U.S. Department of Agriculture (USDA), and the Environmental Protection Agency (EPA). The task force is thus a truly interagency body, drawing on a variety of agency expertise for leadership.

The legislative history describes the task force as being charged with preparing a comprehensive research plan, to include individual research, economic assessment, Federal coordination, international cooperation, and management requirements. The comprehensive plan is implemented and managed by the task force. The Acid Precipitation Task Force could thus serve as a model for an interagency task force dedicated to genome projects.

In 1985, representatives from the various agencies signed a memorandum of understanding that fixed the structure for administering the act. The memorandum assigns authority and responsibility to: 1) a Joint Chairs Council, consisting of principals from USDA, DOE, EPA, NOAA, the Department of the Interior, and the Council for Environmental Quality, and responsible for approving the annual research program and the corresponding portions of the budgets of the participating agencies; 2) the task force, to review the annual research program and budget and to provide advice and recommendations to the council; 3) the Director of Research (appointed by the Joint Chairs Council), to formulate the research program and budget; 4) the Interagency Scientific Committee and the Interagency Policy Committee, consisting of senior scientific and policy executives, respectively, from the agencies, to advise and recommend; 5) an External Scientific Review Panel; 6) the Office of the Director of Research, consisting of scientists and support staff; and 7) research task groups, each under the lead of a specific agency, to develop a research plan and budget for a particular task.

Shortly after the 1985 reorganization, the General Accounting Office reviewed the program at the request of Congress, because management changes and delays in reporting had become constant. The General Accounting Office's recommendations are more functional than structural, and they relate to the difficulty of issuing public reports under great scientific uncertainty. The almost intractable nature of some of the acid precipitation problems is apparently issue-specific and not related to the organization of the project.

The authority under the new organization is significantly vested in the Joint Chairs Council, the Director of Research, and divided between a scientific column and a policy column. The fine structure is fascinating: It attempts to permit each participating agency to retain authority over its research expertise by creating research task groups. For example, Interior is responsible for monitoring deposition, NOAA for atmospheric processes, and DOE for emissions and control technology. In genome projects, distribution according to expertise would have NIH focus on mapping techniques and biological technologies, and DOE focus on automation and robotics and computation. This could be useful as long as it did not assign tasks to the wrong agency and did not inhibit flexible interagency planning for areas of legitimate overlap. The agencies participating in the Acid Precipitation Task Force are working on a scale similar in magnitude to that of a genome project; from fiscal years 1982 to 1987, the agencies spent just over $300 million for acid precipitation research.

The joint chair arrangement, among NIH, DOE, and NSF in a genome project, would represent a smooth distribution of authority. The appointment of a director of research might prove the only bone of contention, as the selection might imply what style of research—small-group science v. Big Science—is to be funded. The Acid Precipitation Task Force also balances the concerns of policy specialists with those of scientists and seeks the input of nonagency scientists as well (it does neglect nonagency policy specialists, however). This interagency task force approach attempts to combine the dynamic properties of an authoritative leader with the efficiency of agencies pursuing their own research expertise.

SOURCE: Office of Technology Assessment, 1987, based in part on U.S. Congress, General Accounting Office, *Acid Rain: Delays and Management Changes in the Federal Research Program*, GAO Pub. RCED-87-89 (Washington, DC: GAO, 1987).

be achieved. A task force or lead agency could eliminate some of this cost.

A task force may not be able to match efficiency in decision making with efficiency in administering the agencies' resources. Its recommendations could be ignored by agencies, or it could prove an obstruction or source of delays. A task force is a bureaucratic solution, identifying a person or group with the goal of genome analysis and building upon the existing authority structure. Such authority may be necessary to direct research and provide a focus for international communications, but it adds another layer to the bureaucracy that separates the administrators of science from the investigators.

Consortium

Like the task force approach, a consortium would involve the creation of a new authoritative entity. Unlike the organizational structures discussed previously, however, this approach would require the active participation of private firms (see figure 6-4). The introduction of this new factor complicates the staging of a genome project.

The typical consortium is a close working association between a university research group and one or more private firms interested in the pursuit and economic development of that research. Government involvement in consortia is often limited to financial support during the initial stages of basic research, while industry waits to fund the development stage. State government is frequently more active than Federal, because the projects are perceived to be closely linked to local economic development (see box 6-B).

Figure 6-4.—Consortium

A consortium of universities, businesses, and government is directed toward several mutually enriching goals: strengthening universities, stimulating (competitive) economic growth, engaging in basic research, creating generic technologies, and developing and delivering specific products (6). These goals correspond to those of some genome projects, which some persons hope will maintain America's competitive position in biotechnology against challenges from Asia and Europe. Genome projects would, for example, seek both to create generic biotechnology tools (such as techniques for handling very large DNA fragments, detecting very small amounts of DNA, and designing software for analysis) and to develop specific products (such as vectors for cloning DNA or automated DNA sequencers). Such results would benefit university researchers and corporate investors alike.

The question of setting the research agenda (in other instances the responsibility of the individuals conducting the research and the agencies funding it or of the task force established to oversee it) is complicated by private firms' need to emphasize technology development and not necessarily free inquiry. Profit-seeking firms often have shorter-term goals than are practical for the support of basic research and therefore focus on the development of short-term technologies over long-term ones. Not all industries have a short-term perspective, however: Pharmaceutical firms are accustomed to basic research and long-term payoffs—investments requiring more than a decade to bear fruit.

A private emphasis on development is closely associated with the effective transfer of technology into the marketplace. A consortium would no doubt speed technology transfer to participating firms, but firms might suppress the spread of scientific information to protect their investment (5). The phenomenon of sitting on data is not restricted to industry—academic scientists may delay dissemination of information in order to consolidate results for their own financial or reputational benefit (12)—but proprietary interest will not help the free exchange of data. Thus, the question of proprietary rights versus information in the public domain is a sticky one for genome projects, where a naturally occurring DNA sequence can translate into a multi-million-dollar product. Thus the presence of commercial firms in academia is two-sided: Goals of technology transfer and economic development may be more easily reached, but the control exerted by industry over the planning of the research agenda and the dissemination of results might be too self-interested. Concern that the economic aspirations of private firms might corrupt the atmosphere of academia may be overstated now, since only a few businesses have shown any interest in the genome project; but the possibilities of reaping economic benefits, especially in the current environment of international competitiveness, are likely to attract more private sector involvement in the future.

Box 6-B.—Midwest Plant Biotechnology Consortium

Representing a large number of university, industry, and government partners, the Midwest Plant Biotechnology Consortium is an experiment in basic research and technology transfer among agricultural sectors. Its purpose is to increase the competitiveness of American agriculture and agribusiness through the development of basic plant biotechnology research.

The idea for the consortium began at DOE's Argonne National Laboratory (ANL), which has a historical research interest in photochemistry and photosynthesis. ANL determined that a coordinated program in plant science could contribute to biotechnology applications of interest to both industry and government agencies.

When ANL invited participation from universities and industry, it specified a number of principles that would guide the consortium. The continued importance of both industrial and scientific peer review processes was stressed, and the intellectual property rights were established from the outset. ANL also emphasized the regional nature of the consortium, encouraging the participation of Midwest institutions to investigate for Midwest agribusiness. Aside from these initial guidelines, the original organization of the consortium remained informal until recently, when it sought incorporation as a 501(c)(3) (tax-exempt) corporation and developed a more formal budget process.

Government interest in the consortium comes from those agencies involved in the genome discussion—DOE, NSF, and NIH—with the addition of USDA. A secretariat operates the consortium, determining policy and procedure with informal involvement of government officials. More formal arrangements may be possible in the future. An executive board of corporate and university officers oversees the technical and administrative operations. A number of research topic subgroups (e.g., plant growth, pesticides-herbicides) also exists, and the primary interaction for technology transfer occurs at this level.

The consortium solves the problem of research direction by a two-tiered system: The industrial partners first select proposals on the basis of commercial potential, and then a peer review system selects on the basis of technical merit. The consortium expects the Federal Government (with some State funds) to support research through the initial stages, that is, until the industrial partners can see around the development corner to a commercial application. Research proposals developed as part of the consortium will be subjected to normal competitive grant review at the Federal agencies. Industry would then fund the final steps.

Organizationally, the Midwest Plant Biotechnology Consortium offers a number of useful parallels to a genome project. The Federal agencies overlap similarly, as does DOE's attempt to link the research program to related research at the national laboratories. Intellectual property rights are sensitive in both projects; the consortium provided from the outset that each research participant would retain rights according to institutional policy and that the industrial participants would have the right to first disclosure. The consortium retains the integrity of the peer review system while allowing industry to set some of its own research priorities based on commercial potential. The parallels break down where some of the short-term commercial interest in a genome project focuses on automated tools and machinery in addition to the results of biotechnical manipulation. Funding for the consortium is also considerably less than what is expected to be necessary for a genome project.

SOURCE: Office of Technology Assessment, 1987.

If consortia related to one or more genome projects are formed, several issues will have to be resolved. First, terms of participation must enable a broad spectrum of private firms to participate. Small firms with limited resources have had difficulty, for example, in paying entry fees to some biotechnology consortia (22). And nonprofit organizations, which must make their information available on a nondiscriminatory basis under U.S. tax laws, might have difficulty in participating if there are preferential terms for industrial partners.

Discussion

None of the four workable administrative structures—the lead agency approach, which requires a choice between DOE and NIH; the cooperative approach, which requires no new legislation; the task force, which creates a formal authority; or the consortium, which adds a dose of private sector assistance—is static. Administrative forms may overlap: For example, the consortium may require a lead agency, or the cooperative effort may create consortia or task forces to attain specific objectives. The administrative structure at the national level does require explicit choices, however. Congressional action will vary according to the option chosen. Interagency agreement and consultation would require no new legislation, only oversight. Designation of a lead agency, establishment of a task force, or creation of a single national consortium would require new legislation.

Administration of genome projects will require monitoring of some central services and facilities, some services and functions performed at centers, and many grants to small groups. This raises several concerns about communication among agencies and among the scientists whose work they support. The diffusion of research among a large number of groups complicates communication, but it permits the most flexible organization of research; the investigator may be as focused or interdisciplinary as the research demands. A reduction in the number of groups reduces the difficulty of communication but limits the number of people trying wholly new approaches to the scientific or technical objective. The pace of innovation may be directly proportional to the number of groups: A commitment to a single center or institute might fix the relevant technology prematurely. When innovation is less important than production, then specialized facilities are logical because they simplify the organization of work. Problems of communication for centrally administered projects are of a different variety. Often the most difficult problem is ensuring that services are appropriate and tailored to the needs of those using them. Different genome projects will have different modes of communication. Projects that rely on many small groups will need communication networks or frequent meetings of scientists; central services will require feedback from user communities.

ADVISORY STRUCTURE

Second to the administrative structure in organizational hierarchy, though not in importance, is the structure of an appropriate advisory body or bodies. Agencies supporting genome projects will benefit from tapping the academic and industrial sectors for the requisite expert wisdom. Similarly, academia and industry wish to ensure their input into the decision-making process and to exercise some control over the research that affects their livelihood. The responsibilities, composition, structure, and funding of advisory groups then become issues.

Responsibilities

The primary responsibility of the independent advisory board (or boards) would be to follow the research plan and budget envisioned by the agencies, task force, or consortium and to make recommendations where appropriate. Such recommendations might include identification of promising research initiatives in need of funding or oversight of standards necessary to ensure quality control. The board could be granted budget authority to enact these recommendations, or its role could be strictly advisory. Consideration of broad overarching issues—such as the ethical implications of using some newly developed technologies or the economic benefits of targeted technology development—could also be a function of the board.

The advisory board would naturally have a reporting duty: to the participating agencies, to Congress, to the public, and perhaps to the international community of scientists. The advisory board would be an organ of communication among the agencies, supplementing their informal direct contact. Congress would probably want to be kept abreast of research progress and could require periodic assessments in order to plan genome projects and other research initiatives. Annual or biannual reporting to Congress on progress and the distribution of funds could be fit into the budget process, for this will be one way in which genome projects are held accountable to the taxpayers. The executive branch could be kept up to date by the advisory board or through the Office of Science and Technology Policy. An advisory board not composed entirely of Federal officers would fall under the Federal Advisory Committee Act (Public Law 92-463). Pursuant to the act, the advisory board's meetings and papers must be open to the public. The advisory board could also be the contact for international communication.

Composition

An advisory board would require members with varied backgrounds. Scientists with experience in the planning of mapping and sequencing work would be needed for technical advice. Scientists with database expertise would also be required, as the storage and dissemination of the project's information is as central as the generation of it. Scientists could be chosen from universities, industry, and federally supported laboratories. Choosing the board involves the same issues as the consortium decision: how much influence development- and profit-minded industry experts should have on the project. One suggestion, from an industrial association, is to set up an advisory board with 50 percent university, 30 percent government, and 20 percent industry representatives (9). This would in fact be an extension of current practice, as university and industry representatives often work together productively. The selection of scientists from abroad to serve on the advisory board, perhaps as nonvoting members, would help it assume an international role.

Since the project's impact would extend into general science policy, economic competitiveness, medical care delivery, and the like, experts from such fields might be included. The board might want, for example, to ensure that other areas of biomedical research do not suffer from a drain of funds or personnel, and policy experts and economists would be helpful in this. Lawyers might be necessary to address questions of intellectual property. Ethicists might be included to help the board address such issues as confidentiality of data on research subjects or whether to investigate the chromosome containing disease gene A before that containing disease gene B. Representatives of interested private philanthropies, particularly those supporting research in human genetics, might also be included. An advisory board would logically include at least a representative of the Howard Hughes Medical Institute, as it funds a substantial portion of genome projects.

Structure and Funding

Scientists and nonscientists could serve together on a single advisory body or on separate bodies. The choice will influence the method of research planning and science policy formation: In a single body, the procedure is multifaceted but essentially unitary; in separate bodies, the procedure is separated into scientific and policy components. Another possible division of advisors would be government representatives on one panel and private representatives, from academia, industry, and other backgrounds, on another.

Appointments to a policy board could be made by the President, with the advice and consent of the Senate. The choice of members could be assigned to a nongovernmental body, such as the National Academy of Sciences, to ensure the board's independence and its technical competence. As an alternative, the task of selection could be delegated to the Office of Technology Assessment.

BIG SCIENCE v. SMALL-GROUP SCIENCE

The likelihood that Big Science will invade molecular biology has often been cited in opposition to a concerted government program of genome projects. Small science is largely conceived and executed by a principal investigator directing a small laboratory group funded by a grant. Big Science can refer to many things. It can mean large and expensive facilities. It can refer to large, multidisciplinary team efforts that entail cooperative planning and therefore require individual scientists to sacrifice some freedom in choosing goals and methods. Or it can refer to bureaucratic central management by government administrators. These different meanings have been intermingled in the emotionally charged debate about genome projects. (For further insight into that debate, see box 6-C.)

Three lines of argument have been made against conducting molecular biology research on one of these Big Science models: style, efficiency, and political interference.

Displacement of Higher-Priority Science

Some scientists worry that a major Federal program to map the human genome and sequence a significant portion of it would detract from the conduct of more important science (2,3,20). The argument is that special appropriations for human genome projects could well go to projects that do not present the most immediate obstacles to scientific progress and might supplant funds that would be allocated differently by the peer review processes of scientific agencies. If genome projects were not of the same scientific caliber as projects in other areas of science, agencies would nonetheless be precluded from reassigning those funds.

Other scientists argue that some genome projects do not lend themselves easily to current review procedures and merit a special effort (7,10, 19,25,30). Genome projects will involve not only science, they say, but also technology development and production. Some aver that existing peer review committees give short shrift to projects intended to develop methodology (as opposed to answering a scientific question) and tend to underfund shared research resources. They believe that the value of genome projects warrants a special effort, including new peer review committees and increased resources for a research infrastructure.

A related issue concerns the details of funding mechanisms. Those who believe strongly in the superiority of investigator-initiated small-group research urge caution in supporting large projects that are administered by institutions rather than individuals. The agencies most directly involved—namely, NIH and DOE—are adopting policies that answer both arguments by promising to use a system of peer review that gives the scientific community substantial power to direct genome projects but that differs from current peer review by adding new review groups to focus on component genome projects.

Style and Efficiency

Some scientists have objected to a Big Science approach to genome projects because it goes against the tradition of science as a cottage industry conducted by small, largely autonomous groups. The underlying assumption is that Big Science management would undercut the motivation and circumscribe the freedom of investigators by making them beholden to administrators in a scientific bureaucracy. Yet team effort is likely to be cheaper and faster in the long run for genome projects that focus on developing instruments or producing maps. It would be unwise and wasteful to shun all projects that do not conform to the small-group mode. One science administrator advised scientists that:

> . . . insofar as what they do is part of the war against human suffering, their desires and tastes are not all that matter. Biomedical science is not done, or, more important, is not supported by the public, simply because it gives intense satisfaction to the dedicated and successful biomedical researcher (32).

Large and expensive projects must meet certain criteria, otherwise they could indeed supplant other research. They must meet needs that cannot be met by small-group research (e.g., produc-

Box 6-C.—Quotes on Genome Controversies

Proposals for genome projects, particularly sequencing the human genome, have provoked considerable controversy among luminaries in molecular biology and related disciplines. The following quotations illustrate the liveliness of the debate over the past 2 years.

"Sequencing the human genome is like pursuing the holy grail." Walter Gilbert, Harvard University, at several national meetings, March 1986 to August 1987.

"[Sequencing the genome now] is like Lewis and Clark going to the Pacific one millimeter at a time. If they had done that, they would still be looking." David Botstein, Whitehead Institute, Cold Spring Harbor Symposium on the Molecular Biology of *Homo sapiens*, June 1986.

"Humans deserve a genetic linkage map. It is part of the description of *Homo sapiens*." Raymond White, Howard Hughes Medical Institute, University of Utah, in *Science* 233:158, 1986.

"The idea is gaining momentum. I shiver at the thought." David Baltimore, Director, Whitehead Institute, in *Science* 232:1600, 1986.

"Of course we are interested in having the sequence, but the important question is the route we take to getting it." Maxine Singer, Director, Carnegie Institution of Washington, in *Science* 232:1600, 1986.

"Sequencing the human genome would be about as useful as translating the complete works of Shakespeare into cuneiform, but not quite as feasible or as easy to interpret." James Walsh, University of Arizona, and Jon Marks, University of California, Davis, in *Nature* 322:590, 1986.

"I believe such a conclusion [against special efforts to sequence the human genome] represents a failure of vision, an unwarranted fear of (not very) 'big' science." Robert Sinsheimer, University of California, Santa Cruz, in *Science* 233:1246, 1986.

"My plea is simply that we think about this project in light of what we already know about eukaryotic genetics and not set in motion a scientifically ill-advised Juggernaut." Joseph Gall, Carnegie Institution of Washington, in *Science* 233:1368, 1986.

"Too bad that it needs such fancy wrappings to attract public attention for an obvious good." Joshua Lederberg, "The Gift Wrapped Gene," in *The Scientist*, Nov. 17, 1986, p. 12.

"The sequence will give us a new window into human biology." Renatto Dulbecco, Salk Institute, interview with OTA staff member, January 1987.

"Of course, if you have the clones, you're going to want to sequence them. The question is which ones to do first. I think it is scientifically arrogant to prejudge what will be important and what will not." Paul Berg, Stanford University, interview with OTA staff member, January 1987.

"I'm surprised consenting adults have been caught in public talking about it [sequencing the genome] . . . it makes no sense." Robert Weinberg, Whitehead Institute, in *The New Scientist*, Mar. 5, 1987, p. 35.

"The sequence of the human genome would be perhaps the most powerful tool ever developed to explore the mysteries of human development and disease." Leroy Hood and Lloyd Smith, California Institute of Technology, in *Issues in Science and Technology* 3:37, 1987.

"The main reason that research in other species is so strongly supported by Congress is its applicability to human beings. Therefore, the obvious answer as to whether the human genome should be sequenced is 'Yes. Why do you ask?' " Daniel Koshland, Editor, *Science* 236:505, 1987.

"The real problem that faces us is not the cost of the Human Genome Program, but how to get it going, seeing both that the right people are in charge and that they work under an administrative umbrella that will not tolerate uncritical thinking and so will never promise more than the facts warrant." James D. Watson, *Director's Report*, Cold Spring Harbor Laboratories, September 1987.

"We will see a new dawn of understanding about evolution and human origins, and totally new approaches to old scientific questions." Allan Wilson, University of California, Berkeley, at a symposium for the director, National Institutes of Health, Nov. 3, 1987.

tion, service, or targeted technology development). They must not merely be useful, but fill critical resource gaps as well (4). These criteria are likely to be met by many databases, repositories, and mapping projects. They have not yet been met by proposals to sequence the entire human genome.

Some argue that, while it may appear that certain projects are best conducted by large, multidisciplinary teams, in the long run science progresses faster if large, targeted projects are not begun (20). That is, small-group science is so much more productive in the long run that attempts to direct science will inevitably go astray.

Similar debates preceded the approval of costly projects in other fields. Construction of cyclotrons and other particle accelerators was resisted by many physicists in the 1930s [Heilbron and Kevles, see app. A], and space-based instruments were opposed by many astronomers in the 1960s (26). Yet these facilities permitted scientific advances that would otherwise have been impossible, and they were (and are) most often used by small research groups. The issue is not that expensive facilities should not be built, but that they should address critical needs and be carefully planned.

Politicization

One way in which concerted projects are believed to drift into inefficiency is through political interference. This can be on a small scale (haggling that impedes progress among members of a research team) or a large scale (e.g., pork barrel science at the national level). One scientist has observed, "a megaproject like sequencing the human genome is certain to increase the political control over scientific decisionmaking" (3), and the American Society for Biochemistry and Molecular Biology warns against "the establishment of one or a few large centers that are designed to map and/or sequence the human genome" (1). Large research institutions can drift once their missions have been accomplished, and it can be difficult to close down unproductive efforts (32).

Molecular biology has been remarkably productive for three decades without the management style of Big Science. In the recent inventory of 275 Big Science facilities compiled by the House Committee on Science and Technology, none was biological (29). Yet some human genome projects, for example developing new instruments or pooling results from many different groups, will require multidisciplinary teams concentrating on a technical problem. This situation is analogous in many ways to the situations faced earlier by other sciences in their transition to Big Science [Heilbron and Kevles, see app. A] (32). It is difficult to imagine, for example, automating the steps in cloning DNA, sequencing it, or mapping it without combining optics, chemistry, physics, engineering, and electronics. If the end products of genome projects—materials and information—are to be reliable and used internationally, there must be quality control and standardization.

Clearly, some important functions require central coordination or multidisciplinary team research, although not necessarily centralized administration; some tasks cannot be forced into the mold of small-group science. Technological developments will determine the pace and extent to which Big Science becomes part of biological research. The question will be how to decide which projects merit special effort and which do not. Decisions of several types will be necessary in conducting genome projects. The advantages of decentralized planning must be balanced against the need for some centralized resources. The importance of mapping, sequencing, and technology development must be compared to other research and services. Such decisions will require an administrative structure to make them.

Biomedical investigations are now, and in the foreseeable future will continue to be, conducted primarily by small groups, although Big Science facilities and services can amplify and complement them. Small groups will remain the principal means of studying physiology and disease. When new institutions are created for elements of human genome projects, special attention must be paid to making results useful to small scientific groups. It would be ironic if genome projects starved small-group research efforts in order to create new tools.

The costs of database, repository, and map projects are not large relative to the costs of other biomedical research, so planned projects are unlikely to have any measurable adverse impact on other research. Moreover, genome projects intended to bolster the research infrastructure should free funds for new work by making research faster and less costly. If genome projects threaten the health of small-group biomedical research, then genome projects should take a back seat.

SUMMARY

The Howard Hughes Medical Institute recently issued a short report on efforts to map the human genome; it observed:

> The sooner the entire genome is mapped and sequenced once and for all, the sooner scientists can get on with the real work of human biology: understanding what the genes do (21).

Databases and repositories must be centrally administered, although not necessarily centrally located, in order to be widely accessible. Technology will most likely determine whether and when large facilities and coordinated administration are necessary to conduct genome projects. If large facilities prove to be more efficient, this will not necessarily be incompatible with research by small groups; it could in fact enhance it. If, however, large facilities and centrally organized research programs threaten the lifeblood of biomedical research—investigator-initiated grants—then the projects should be reevaluated and, if necessary, cut back.

CHAPTER 6 REFERENCES

1. American Society for Biochemistry and Molecular Biology, policy statement, April 1987.
2. Ayala, F., "Two Frontiers of Human Biology: What the Sequence Won't Tell Us," *Issues in Science and Technology* 3:51-56, 1987.
3. Baltimore, D., "Genome Sequencing: A Small-Scale Approach," *Issues in Science and Technology* 3:48-50, 1987.
4. Baltimore, D., Whitehead Institute, personal communication, December 1987.
5. Cahill, G., comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
6. Dimancescu, D., and Botkin, J., *The New Alliance: America's R&D Consortia* (Cambridge, MA: Ballinger, 1986).
7. Gilbert, W., comments at Costs of Human Genome Projects, OTA workshop, Aug. 7, 1987.
8. Gilbert W., "Genome Sequencing: Creating a New Biology for the Twenty-First Century," *Issues in Science and Technology* 3:26-35, 1987.
9. Godown, R.D., testimony on Title II of S.1480 before the Subcommittee on Energy Research and Development, Senate Energy and Natural Resources Committee, Sept. 17, 1987.
10. Hood, L., and Smith, L., "Genome Sequencing: How To Proceed," *Issues in Science and Technology* 3:36-46, 1987.
11. Joyce, C., "The Race To Map the Human Genome," *New Scientist*, Mar. 5, 1987, pp. 35-40.
12. Karny, G., comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
13. Kirschstein, R.L., letter in *Issues in Science and Technology*, summer 1987, p. 5.
14. Koshland, D.E., "Sequencing the Human Genome," *Science* 236:505, 1987.
15. Levinson, R., NIH, personal communication, June 1987.
16. Lewin, R., News and Comment, *Science* 235:1453, 1987.
17. Mitra, S., Oak Ridge National Laboratory, personal communication, March 1986.
18. National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
19. Office of Technology Assessment interviews with R. Dulbecco, L. Smith, T. Friedmann, M. Bitensky, R. Davis, and P. Berg, January 1987; W. Gilbert, July 1987.
20. Office of Technology Assessment interviews with A. Kornberg and S. Cohen, January 1987; D. Baltimore, July 1987.
21. Pines, M., *Mapping the Human Genome* (Bethesda, MD: Howard Hughes Medical Institute, 1987).
22. Raines, L., Industrial Biotechnology Association, personal communication, November 1987.
23. Roberts, L., "Agencies Vie Over Human Genome Project," *Science* 237:486-487, 1987.
24. Science and the Citizen, "Geneshot," *Scientific*

American, May 1987, pp. 58-58a.

25. Smith, L., and Hood, L., "Mapping and Sequencing the Human Genome: How To Proceed," *Bio/Technology* 5:933-939, 1987.
26. Smith, R., The Johns Hopkins University, personal communication, December 1987.
27. Subcommittee on the Human Genome, Health and Environmental Research Advisory Committee, *Report on the Human Genome Initiative*, prepared for the Office of Health and Environmental Research, Office of Energy Research, Department of Energy (Germantown, MD: DOE, April 1987).
28. U.S. Congress, General Accounting Office, *University Funding: Information on the Role of Peer Review at NSF and NIH*, GAO Pub. RCED-87-87FS (Washington, DC: U.S. Government Printing Office, March 1987).
29. U.S. Congress, House Committee on Science and Technology, *World Inventory of "Big Science" Instruments and Facilities* (Washington, DC: U.S. Government Printing Office, 1986), cited in Heilbron and Kevles, see app. A.
30. Watson, J.D., director's report, Cold Spring Harbor Laboratories, in press.
31. Watson, J.D., comments at Costs of Human Genome Projects, OTA workshop, Aug. 7, 1976.
32. Weinberg, A.M., *Reflections on Big Science* (Cambridge, MA: MIT Press, 1967), esp. pp. 113-114.
33. Zinder, N.D., report of the ad hoc committee reviewing the Special Virus Cancer Program of the National Cancer Institute, March 1974.

Chapter 7

International Efforts

CONTENTS

Page
Introduction 133
Japan 136
  Mapping and Sequencing Research 136
  Potentials for Cooperation and Conflict With the United States 138
Europe 139
  European Organizations 139
  National Research Efforts in Europe 143
Other International Efforts 148
  Australia 148
  Canada 148
  Latin America 149
  South Africa 149
  The Union of Soviet Socialist Republics and Eastern Europe 149
International Collaboration and Cooperation 150
  Precedents for International Scientific Programs 150
  Options for International Organization of Genome Research 152
  Existing Collaborative Frameworks 155
Chapter 7 References 159

Boxes

Box *Page*
7-A. The Venezuelan Pedigree Project 134
7-B. The Center for the Study of Human Polymorphism (CEPH): An International Gene Mapping Center 146
7-C. The International Geophysical Year 151
7-D. Views on International Cooperation and Collaboration in Genome Research 152
7-E. Large Centers v. Networking 156

Figures

Figure *Page*
7-1. Distribution of Publications in Human Gene Mapping and Sequencing 133
7-2. Human Gene Mapping and Sequencing Articles Published Annually 158

Table

Table *Page*
7-1. International Distribution of Human Genome Research 133

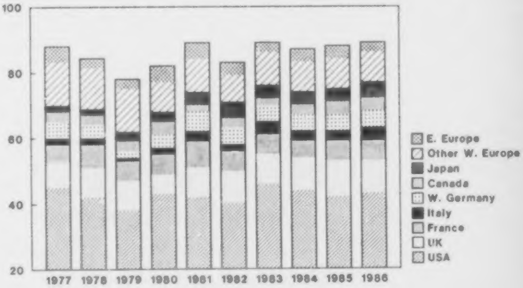
Chapter 7

International Efforts

INTRODUCTION

The expected health benefits of genome projects—and their commercial potential—have attracted international as well as national attention. The United States is the clear leader in basic research, publishing more articles on mapping and sequencing than European or Asian nations (see figure 7-1, table 7-1). U.S. companies have also marketed more instruments for DNA research than any others (see ch. 2). Productivity in basic and applied research does not, however, guarantee the United States the lead in developing or producing commercial products and processes, nor does it ensure market competitiveness. Japan has also encouraged the commercial development of technologies associated with the mapping and sequencing of DNA. Countries such as Switzerland and West Germany are home base for multinational pharmaceutical and chemical companies that are poised to commercialize developing products. Some nations not supporting much basic genome research at present have strong biotechnology or high-technology resources and policies and might

Figure 7-1.—Distribution of Publications in Human Gene Mapping and Sequencing



Compiled from a bibliometric analysis of literature on human gene mapping and sequencing conducted for the Office of Technology Assessment by Computer Horizons, Inc. [see apps. A and E]. The differences between the annual percentages displayed and the total annual research (100%) can be attributed either to countries not included in the listing or to the absence of sufficient bibliographic information to determine the country or region from which the publication originated.

SOURCE: Office of Technology Assessment, 1988.

Table 7-1.—International Distribution of Human Genome Research (percent of articles published annually on human gene maps or markers)

Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
United States	45%	42%	38%	43%	42%	40%	46%	44%	42%	43%
Japan	2	2	3	3	4	5	4	4	4	5
Western Europe										
Denmark	1	3	1	1	2	1	1	1	1	1
Federal Republic of Germany	5	4	2	4	6	5	5	5	5	5
Finland	<1	1	1	1	1	<1	<1	1	1	<1
France	5	7	6	6	8	6	6	5	6	6
Italy	2	2	1	2	3	2	4	3	3	4
Netherlands	4	5	3	3	2	2	2	2	3	2
United Kingdom	8	9	9	6	9	10	9	10	11	10
Other	7	4	7	4	6	5	6	5	4	5
Other non-European countries										
Australia	<1	1	2	2	2	2	2	2	1	2
Canada	3	3	3	4	2	3	2	3	4	4
Eastern Europe and U.S.S.R.	5	3	3	5	5	4	3	4	4	3
South Africa	0	1	1	1	<1	<1	<1	<1	1	<1
Other	5	4	6	4	4	5	3	3	5	3

SOURCE: Office of Technology Assessment, 1988, compiled from a literature search and bibliometric analysis conducted for the OTA by Computer Horizons, Inc. The key words used in the search are described in app. E.

be well positioned to commercialize technologies that are developed for and spun off from human genome research.[1] The OTA has found that Government agencies in the United States are further along in developing policies for genome projects than are comparable agencies in other countries, although a number of other countries have well-established basic research efforts in mapping and sequencing human and nonhuman genomes—efforts that could either complement or compete with U.S. efforts.

Gene mapping is perhaps the most common international research activity in human genetics, and it is likely to be an area to which many nations will contribute. Human genes are highly polymorphic, and populations from different regions exhibit considerable genetic variation. These regional differences will allow researchers to contribute to comparative studies, as well as to characterize and map genes of particular regional interest (e.g., the thalassemias in the Mediterranean and Oudtshoorn skin disease in the Afrikaner population in South Africa). The study of DNA from diverse peoples will shed light on the nature of polymorphisms and genetic disorders, even if it does not lead immediately to improved health care (8).

The large scope of genome projects invites international cooperation. Informal cooperation and collaboration are already underway through a variety of mechanisms. Formal collaboration could speed research and reduce the financial burden on each country. **Maintenance of international databases and repositories is particularly important to provide timely access to information from research conducted around the world**. Many scientists encourage international cooperation in genome research, but any effort to conduct genome mapping and sequencing projects on an international scale must be based on a realistic assessment of the capabilities and interests of the countries involved.

Countries that do not themselves carry out the kinds of research involved in mapping and sequencing can play an important role by collecting genetic material from families for comparative studies. One such project, a collection of genetic material from a group of Venezuelan families, was a key factor in the successful search for the gene that causes Huntington's disease (see box 7-A). Similar pedigree collections are being established and maintained in Egypt and Denmark, as well as in isolated populations in the United States such as Mormon and Amish communities. These pedigrees provide valuable source material for the study of polymorphisms and genetic disease in human beings.

This chapter summarizes the state of DNA mapping and sequencing research in Japan, Western Europe, and elsewhere. Issues of international cooperation and competition and precedents for international cooperation in science are examined. Some organizational options for the international management of genome projects are proposed, specifying areas in which cooperation might best be achieved and describing cooperative frameworks already in existence. Chapter 8 outlines questions about international technology transfer that might emerge in collaborative or cooperative situations.

---

[1] It is not within the scope of this assessment to provide a detailed analysis of biotechnological capabilities and industrial funding; suffice it to say that genome research is part of a much larger arena of Federal, university, and industrial research and development. A forthcoming OTA assessment, *New Developments in Biotechnology, 4: U.S. Investment in Biotechnology* (81), covers this in great detail for the United States. A previous OTA assessment, *Commercial Biotechnology: An International Analysis* (79), describes the state of biotechnology in Western Europe and Japan; the more recent Department of Commerce reports, *Biotechnology in Western Europe* (91) and *Biotechnology in Japan* (39), offer updated information on international efforts.

Box 7-A.—The Venezuelan Pedigree Project

In the small fishing villages that line the coast of Lake Maracaibo in Venezuela lives an unusual group of families. If you walk into any of these villages, you may be met by residents who do a characteristic dance down the streets—large, jerky motions, staggering and weaving from side to side. For many years the residents of these villages were ostracized, considered to be chronically drunk. But in the early 1970s a doctor from a nearby military base realized that the dance was not due to alcoholism but to Huntington's disease, a rare, dominant genetic disease that causes degeneration of nerve cells in the brain. The onset of Huntington's disease is generally late: In those who carry the gene, symptoms begin at age 35 or older. The disease leads to loss of control of the voluntary muscles, first causing twitches and jerks, then dementia, and finally death.

A preliminary study describing the case histories and pedigrees of approximately 100 patients from Lake Maracaibo families was presented at a meeting of the American Neurological Association in 1972. It was an interesting case: an interrelated set of families, along whose pedigree could be traced an extraordinarily high incidence of a genetic disease that is rare in the general population. At the time, however, no one knew what to do about it. The case remained an interesting anecdote in the memories of the researchers who attended the meeting.

One of those researchers was Nancy Wexler, a clinical psychologist. Wexler had both a professional and a personal interest in Huntington's—her mother had died of the disease, so she and her sister each have a 50:50 chance of developing it.

Wexler and her colleagues remembered the case of the Venezuelan families 5 years later, when writing a report on Huntington's disease for a congressional commission. One of their recommendations was to initiate a genetic study of the Venezuelan pedigree. Starting in 1979, the National Institutes of Health appointed Wexler to direct a program that would implement the recommendations and set aside funding for the Venezuelan genetic study. The first team of researchers went to Lake Maracaibo in 1981 to collect blood samples from which to extract DNA. At the same time, they compiled a careful record of the pedigrees of the volunteers from whom the blood was extracted. Research teams have gone every year since. The pedigree has grown to include over 7,000 family members; the diagram of it occupies a 100-foot-long section of a corridor near Wexler's Columbia University office. DNA samples have been collected from nearly 1,500 family members, some with Huntington's and some without.

At the same time the genetic study got underway, advances in recombinant DNA technology, specifically the elaboration of techniques for finding genetic markers using RFLPs (see ch. 2), increased the power of analytical methods that could be used on the collected family materials. In 1982, Jim Gusella and others began to screen the DNA from the Venezuelan collection for genetic markers linked to the gene for Huntington's disease. They tested DNA from normal and affected members of the Venezuelan families, comparing the different patterns cut by restriction enzymes on the samples from different family members. The fact that the pedigree included large extended families was useful in locating informative markers. By 1983, the researchers had figured out which chromosome contains the Huntington's gene and had identified a linked marker, paving the way for a diagnostic test—an extraordinary breakthrough, says Wexler, in such a short time. The search for the actual gene is not yet over, however, since locating more closely linked markers has presented unforeseen difficulties.

A cure is not in sight for the families of Lake Maracaibo, but they have made an extremely important contribution to the study of Huntington's.

Photo credit: Nick Kelsh/Kelsh-Marr Studios

Huntington's patient being rowed across Lake Maracaibo.

Photo credit: Frank Micelotta/Time magazine

Nancy Wexler going over Huntington's disease pedigrees.

Moreover, their genetic materials are a valuable resource for other genetic studies, including searches for other disease genes, as well as for the development of genetic maps. Indeed, some of the DNA has been contributed to the international mapping collaboration coordinated by CEPH (see box 7-B). Wexler suggests that "the pedigree is a big genetic playground—whatever idea you have, you could probably test it there."

The Venezuelan pedigree project highlights an important role that developing countries can play in human genome projects, even if they do not yet have the capability to carry out human genome research on their own. A similar collection of genetic materials from patients with genetic diseases (primarily the anemias and thalassemias) and their families was started in Egypt in 1964 and has proceeded since in collaboration with scientists from NIH and several universities—Oxford, London, Harvard, Columbia, New York University, and the University of California, San Francisco. The scientists who manage this collection are eager to cooperate in international efforts to map and sequence the human genome. As Wexler points out:

> In many cases, the countries are eager to collaborate, but they don't know what they have to offer. The patient populations are a valuable resource. And once the working relationships are established between Third World countries with health problems and the high-tech labs in the developed countries, the connections are there for advice and assistance if those countries get to the point of starting their own labs.

SOURCES

J. Gusella, "Gene, Mapping and Disorders of the Nervous System," lecture at American Association for the Advancement of Science annual meeting, Boston, Feb. 15, 1988.

N. Hashem, Ain Shams University Medical Center, Cairo, Egypt, personal communication, July 1987.

N. Wexler, Columbia University, personal communication, October 1987.

JAPAN

Japan's efforts to develop automated DNA sequencing technologies have been highly publicized over the past year, causing concern that Japan will capture the market for sequencing technology and that it will realize most of the potential profits from genome projects. Japan does not, however, have well-defined government policies for human genome mapping. Instead, funding for mapping and sequencing research is under the jurisdiction of half-a-dozen government agencies that often compete for prestige rather than attempt to coordinate efforts.

Mapping and Sequencing Research

The general framework for science policy in Japan is formulated by a small group of bureaucrats in the various agencies and by an inner cabinet group, the Council on Science and Technology, chaired by the prime minister. Programs for human genome research have been divided among the Ministry of Education, Science, and Culture (MESC), the Science and Technology Agency (STA), and the Ministry of International Trade and Industry (MITI). The Ministry of Agriculture, Forestry, and Fisheries supports some research on nonhuman genomes, notably a $500,000 feasibility study on sequencing the entire genome of rice (77).

The Ministry of Education, Science, and Culture

Most mapping and sequencing research falls under the domain of MESC, the primary supporter of basic research in Japan. Like the National Institutes of Health in the United States, MESC supports research projects selected by peer review; it provides grants and funds for universities and university-based researchers and for several national research institutes. In addition, the ministry can encourage research in specific, targeted areas on the recommendation of its advisory committees.

The ministry does not yet have an official policy regarding genome research but has appointed an advisory committee to study the situation. Members of the committee visited the United States in early 1988 to gather information on U.S. policies on human genome research and to ascertain what the U.S. expects of Japan. The committee's recommendations will be implemented beginning in fiscal year 1989 or 1990 (58).

Japan is often criticized for not doing enough basic research; many observers have questioned whether Japanese scientists have enough expertise in basic molecular biology to support a major gene mapping or sequencing effort [Yoshikawa, see app. A]. Bibliometric analysis [see app. E] indicates that while Japan's research output in DNA mapping is far below that of the United States (figure 7-1, table 7-1), its proportion of research relative to other countries has consistently increased over the last decade. Its share of publications on human gene mapping and sequencing rose from 2 percent in 1977 to 5 percent in 1986, compared to a U.S. share that varied from 40 to 46 percent during those years. In addition, MESC supported the research of a scientist that led to the publication of a complete genetic map of *E. coli* in the prestigious magazine *Cell* in 1987 (53). U.S. researchers published a map of *E. coli* at about the same time, but the Japanese research was notable for the speed with which it was done and for the use of automated technologies.

The Science and Technology Agency

STA supports mostly mission-oriented basic research. It has played a leading role in the development of automated sequencing technology. Since 1981, STA's Special Coordination Fund for the Promotion of Science and Technology has underwritten a program entitled Extraction, Analysis, and Synthesis of DNA, with a total funding of $3.8 million (40). The project, led by Akiyoshi Wada of the University of Tokyo, aims to "to reduce the burden of time demanded of researchers working on the analysis of DNA base sequences by developing automatic machinery," utilizing the knowledge and resources of companies with expertise in electronics, robotics, computers, and material science [Wada quoted in Yoshikawa, app. A]. The project scientists are adapting robotic techniques and mass production machines to automate the time-consuming steps in the Maxam and Gilbert sequencing process (see ch. 2) rather than developing new processes. The project has resulted in a prototype of a microchemical robot, made by Seiko, but it is not yet on the market. The goal of the project has been to increase the rate of DNA sequencing output in general, not to sequence the entire human genome. Wada has repeatedly emphasized the necessity for international cooperation in the project and would like to develop a supersequencing center to operate as a service facility for scientific groups around the world (84,87).

STA and a private foundation sponsored an international conference in Okayama in July 1987 to discuss the state of DNA sequencing technologies and possible strategies for genome sequencing in the future; the conference gave no clear indication of the pace or direction of future STA efforts. Some scientists expressed doubts about the STA project, noting that there has been no public discussion in Japan about whether or not to support Wada's conception of the project and that the project is not actively supported by many other Japanese scientists [Yoshikawa, see app. A]. Still, a quiet consensus has emerged that sequencing technology should be developed regardless of whether a full-scale project to sequence the human genome is launched.

Oversight of the project has now shifted from the Special Coordination Fund to STA's Council for Aeronautics, Electronics, and Other Advanced Technologies (CAEOAT); a decision on the status of future directions of the sequencing research should be made by spring 1988. The publicity and momentum of the project are undoubtedly attributable in part to the active role that ex-Prime Minister Nakasone played in advocating biotechnology and related projects [Yoshikawa, see app. A]. Whether the momentum will continue now, after Nakasone's retirement, remains to be seen.

The Ministry of International Trade and Industry and the Human Frontiers Science Program

MITI coordinates applied research, linking university researchers with industry to encourage technology development and commercialization. It does not now play a major role in genome research, but its influence may increase if the Human Frontiers Science Program is fully funded. A human genome sequencing project may become a focal point for the program.

The Human Frontiers Science Program (HFSP) is a proposal for an international, cooperative program of research in basic biology and the development of related "key technologies." The proposal originated in 1985 in MITI's Agency of Industrial Science and Technology (AIST). The proposal came about partly in response to international criticism that Japan does little basic research itself, but capitalizes on the research of others (4,23), and partly to emphasize international cooperation in the face of persistent foreign trade frictions [Yoshikawa, see app. A]. The HFSP proposal met with a lukewarm reception during early outings and international conferences, however, and Nakasone's mention of it at the June 1987 Economic Summit meeting roused little enthusiasm [Yoshikawa, see app. A] (46,76).

If implemented, HFSP would probably enhance Japan's sequencing effort, since DNA sequencing technology has been identified as a key area for development. The program was granted an initial budget of 197 million yen (approximately $1.5 million) for fiscal year 1987, to conduct a feasibility study, but the amount to be spent on development of sequencing technologies is not yet clear. Some observers speculate that the proposal will be shelved now that Nakasone has retired. MITI officials contend, however, that the program is still viable (4,90). A December 1987 planning meeting again endorsed human genome sequencing as a focus for HFSP, but the Ministry of Finance probably will not decide on the program's budget until 1989 (75).

Commercialization of Mapping and Sequencing Technologies

Potentially marketable technologies that are developed for genome projects have been supported by the several mechanisms through which the government aids industrial research in technology development. STA's Special Coordination Fund, established in 1981, provides incentives for basic research for new technologies in accordance with the long-term goals for science and technology development set by its Policy Committee. STA's Research Development Corp. promotes commercial uses of government-developed technologies that might not be used otherwise. The prototype of Seiko's microchemical machine was developed with assistance from the Special Coordination Fund, while the Research Development Corp. has supported its commercialization. In addition, Hitachi, Fuji Photo Film, Toyo Soda, and Mitsui Knowledge Industries have all undertaken research into the automation of DNA sequencing, and some relevant products are being commercialized. DNA extractors developed by Toyo Soda are already on the market, as is a gel preparation by Fuji and autoradiograph readers by Seiko and Hitachi.

Potentials for Cooperation and Conflict With the United States

Many Japanese scientists are willing to cooperate in an international genome sequencing project, but collaboration will clearly be accompanied by economic tensions and competitive posturing both by the United States and by Japan.

The development of similar automated technologies by U.S. and Japanese companies may pose difficult trade issues. The Japanese concentration on sequencing hardware has drawn criticism from American companies, which fear that the Japanese could take the lead in developing technologies for the analysis of DNA (89). At present, however, U.S. manufacturers are clearly ahead in the development and manufacture of equipment for manipulating and analyzing DNA (see ch. 2). Japanese companies are not as far along in marketing relevant products as is often reported—while the Seiko machine has been touted in the Western press, few scientists in Japan have even heard of it (40). In addition, the machine's economy has been overrated: One frequently quoted estimate for the sequencing systems is $0.17 per base pair, with a target of $0.01 or less, but Wada himself states that the system is still far from reaching even the $0.17 goal (86). (The present cost of sequencing is approximately $1.00 per base pair.) Finally, despite the customary preference of Japanese officials for buying Japanese machines, officials of U.S.-based Applied Biosystems, Inc. (ABI, Foster City, CA) in Japan have reported no difficulty in marketing their DNA sequencing machines and other instruments used by molecular biologists [Yoshikawa, see app. A]. To date, Japan is the largest market for ABI's sequencing machine (47).

One frequently voiced fear is that Japanese companies are focusing on automating parts of the sequencing process that companies in the United States have not yet automated (although several U.S. firms have begun development). Thus far, however, the STA-sponsored technology development effort is based on automating machines that use conventional methodology rather than developing or using new molecular biology techniques. Scientists at some U.S. companies have commented that it may have been a mistake for Japan to invest so much in automating existing methodologies when there are new technologies emerging that may make the old methods obsolete.

Databases, which are generally considered useful and politically straightforward areas for cooperation on genome projects, present knotty problems of ownership of information. Despite support within the scientific community, the development of shared databases—even within Japan—is problematic. The Japanese Government has recognized that Japanese databases and repositories are insufficient to handle even its own research and development, and it is trying to establish the database infrastructure necessary for a sequencing effort. It appears, however, that the effort is not well coordinated: Nearly every one of the government agencies is setting up a DNA or protein sequence database for its own purposes, with a minimum of interaction. The DNA Data Bank of Japan (DDBJ), initially established in MESC's National Institute of Genetics in 1984 as a counterpart to GenBank® in the United States and the database operated by the European Molecular Biology Laboratory (EMBL), has lacked adequate staff and computing power. Until recently, it operated only as an access node to GenBank® and EMBL. It has stepped up its operations, however, and is now gathering and entering data from Japanese researchers and transmitting it to the other databases (see app. D). DDBJ formally joined the GenBank®/EMBL collaboration in May 1987; the Japanese data were released in the most recent updates of GenBank® and EMBL.

EUROPE

While Japan is often viewed as a prime competitor, many European countries have stronger research traditions in molecular genetics and the development of related technologies. There are notable genome mapping and sequencing activities in France, Italy, and the United Kingdom, and significant research in gene mapping and technology development in Denmark, the Federal Republic of Germany, and others. In addition, several supranational organizations in Europe have developed targeted programs to encourage biotechnology development; human genome projects can be and are being included. The following sections describe research activities underway in the European community as a whole and in selected countries, in alphabetical order.[2]

European Organizations

Over the past two decades, many European nations have supported scientific collaboration in principle, but in practice funding has been a persistent problem:

> Most European governments have become increasingly reluctant to invest large sums of public money in domestic and civilian R&D, and this is reflected at the European level. . . . As domestic science budgets in Europe have become hard-pressed for cash, governments are asking whether they are getting value for money from international projects. Scientists in some fields have also come to view such projects as unwelcome competitors for their domestic research budgets (29).

Nonetheless, several existing organizations in Europe either support genome research now or could do so in the future.

The European Economic Community

The founding treaties that established the institutions of the European Economic Community (EEC) made little explicit provision for research and development beyond that needed for Euratom (which dealt with nuclear energy, including radiation biology), the Coal and Steel Community, and some coordination of agricultural research under the Treaty of Rome founding the EEC. In January 1974, the Council of Ministers agreed on the general need for an EEC research and development policy, and in the mid-1970s, the EEC's advisory commission began proposing programs, including a program of research and training in selected areas of genetics and enzymology (biomolecular engineering). It was not until 1981 that this proposal was approved, since Article 235 of the Treaty of Rome specifies that such programs can only be adopted by unanimous agreement of all member states (11).

---

[2] The information presented in the sections on selected countries is based on several sources. The OTA contracted a report on research efforts in key countries in Western Europe [Newmark, see app. A]. Some information was gleaned from scientific journals and international news sources. In late 1986 and throughout 1987, OTA conducted an informal survey of international efforts, contacting embassy officials, science attaches, and scientists from numerous countries to request information about the types and funding levels of genome mapping and sequencing research undertaken in those countries and asking whether any specific policies governed genome research. The information gathered from this effort varied considerably in focus, depth, and detail. The countries represented here—other than those with targeted or particularly well known research programs—are thus self-selected and self-reported. The result is a descriptive account rather than a comprehensive analysis.

Support of research and technological development has been enhanced by the adoption of the Single European Act, which took effect on July 1, 1987. This act modifies and extends the Treaty of Rome by adding provisions for precompetitive research to strengthen "the scientific and technology basis of European industry and to encourage it to become more competitive at the international level" (19). Once a multiyear framework program is unanimously agreed on by member states, the individual research and development programs within its agreed areas and financial limits can be approved by a qualified majority (member state votes are weighted roughly by size). The current framework program, an initiative to help create collaboration in targeted areas in science and technology, was adopted on September 28, 1987, and runs until 1991, with a global limit of 5.396 million ECUs (European currency units, which in recent years have had approximately the same value as the U.S. dollar). Framework programs must be proposed by the commission and approved by the governing Council of Ministers and the European Parliament (11).

Most relevant to genome research is a series of research programs in biotechnology: the Biomolecular Engineering Programme, 1981-85; the Biotechnology Action Programme (BAP), 1985-89; and Biotechnology Research and Innovation for Development and Growth in Europe (BRIDGE), 1990-93. A Concertation Unit for Biotechnology in Europe was established in 1984 to coordinate the various activities in biotechnology [Newmark, see app. A]. These programs have been designed to complement national research programs while promoting the development of European biotechnology (83).

The budget for BAP has been substantially reduced from the original proposal; as of spring 1987, it appeared that approximately $300 million of the proposed $6 billion budget would be earmarked for biotechnology research, with another $100 million for health, including some funds for human genome mapping and sequencing work, under the heading of "predictive medicine" [Newmark, see app. A]. "Within the biotechnology program(s), active consideration is being given to mapping and sequencing technology, and in particular with respect to the genome of yeast," although "given the range of topics within the current biotechnology program, it would be surprising if genome work gained more than a small fraction of the total" (11). However, "Community research expenditures have a catalytic role that mobilizes other funds, and a political significance that enhances the coherence and consequent effectiveness with which national funds are deployed" (11). BAP encourages proposals that include at least one industrial partner in the research effort or that provide specific evidence of interest on the part of industry.

When BAP expires, it will be replaced by BRIDGE, which is likely to place even more emphasis on industrial participation. While not yet finalized, BRIDGE is likely to include a project to sequence the genome of yeast, which is more feasible than sequencing the human genome [Newmark, see app. A]. The tentative plan is to undertake a 2-year pilot project in which perhaps 15 laboratories will concentrate on sequencing one yeast chromosome; eventually, a large number of European yeast laboratories would be involved. The pilot project might be launched under BAP, but the full project would be part of BRIDGE and is provisionally estimated to cost $50 million. The project would also try to create a market for sequencing equipment [Newmark, see app. A]. Research on the project will begin soon at some participating laboratories in the United Kingdom [Mount, see app. A].

A subprogram of BAP, Contextual Measures for R&D in Biotechnology, aims to enhance EEC capabilities in bio-informatics (the use of computers and information science in biology), data capture techniques (including advanced instrumentation and automated reading), data banks, computer modeling, computer software, and the "collection of biotic materials" (repositories), along with the "development of information and communication techniques for enhancing the quality and usefulness of such collections" and "the development of techniques for the identification, characterization, conservation, and resuscitation of the materials held in such collections" (20). Development of a biotechnology infrastructure has obvious potential for researchers in human genetics.

Another EEC activity that aids genome research is the Task Force for Biotechnology Information. Created in 1982, the task force has produced discussion papers and has provided small sums of money, totaling $200,000, to support databases (including a contribution to software development at the database of nucleotide sequences run by the EMBL, discussed below and in app. D), and the launching of the European branch of the CODATA Hybridoma Databank, centered at the American Type Culture Collection in Rockville, Maryland. The task force work plan for 1987-90 maintains support for databases, communications, and computational research. The commission of the EEC also supported a series of workshops and studies (1984-86) investigating the interface between biotechnology and information technology in a planning exercise known as Bioinformatics: Collaborative European Programs and Strategy (BICEPS), which "aims to formulate a mid- to long-term strategy for Europe in bio- and medical informatics" and "overall, to improve the European competitive position in the rapidly developing world market for these technologies and applications" (18). Documents for BICEPS refer to the informatics requirements of human genome sequencing and have contributed to plans for bio-informatics in BAP and BRIDGE and to a proposal for a program of Advanced Informatics in Medicine (17). The proposed pilot phase, 1988-90, at 25 million ECUs, was presented by the commission to the European Parliament and the Council of Ministers in September 1987. It includes plans for the development of advanced sequencing instruments and related computational facilities required in genome and other areas of biochemical and protein engineering research. The European chemical industry trade association has endorsed some of the BICEPS proposals and has indicated a willingness to help support an infrastructure such as sequence databases (11,21).

Apart from biotechnology programs, EEC funds research and development in health. The commission's original proposals for the framework program envisaged a Program of Predictive Medicine and Novel Therapy, which would seek "development of predictive medicine and novel therapy oriented towards better knowledge of the human genome, and genetic engineering processes aiming at the repair of DNA defects (e.g., in congenital diseases of genetic origin)" (11). The program was designed to support research in four areas from 1987 through 1991: study of the human genome (including mapping the genome as an aid in the diagnosis and prevention of genetic disease), nucleic acid probes, genetic therapy, and monoclonal antibodies. Funding for the program, originally proposed at $75 million, has been revised downwards to $25 million; both budget and content may be further revised before the program is approved.

The European Molecular Biology Organization

Funded by 17 European countries, EMBO serves primarily to strengthen the training of European molecular biologists. It supports fellowships, workshops and training courses, occasional scientific meetings, and a journal, but it does not directly support research. EMBO sponsored a meeting of Europeans with an interest in human genome research in spring 1987. Few of the scientists present expressed an interest in mounting a major European mapping or sequencing project; instead, most favored informal cooperation between individual laboratories. The group was pessimistic about whether public funds could be found for a large-scale project and raised the possibility of seeking private funds [Newmark, see app. A].

The European Molecular Biology Laboratory

Located in Heidelberg, West Germany, EMBL is financed by contributions from 10 of the 17 member nations of EMBO. It houses the administrative offices of EMBO, but the organizations have separate budgets and purposes. EMBL's staff of about 250 scientists and technicians, drawn from member nations and from West Germany, work on a scientific program proposed by its director-general, at present Lennart Philipson, and subject to the approval of a council composed of representatives from contributing countries. The laboratory was founded with the notion that molecular biology would require facilities that would be too expensive for any national research program to support. For the most part, however, research in molecular biology has not required large centralized facilities, and member nations have tended to interact less with EMBL as they have become proficient at molecular biology in their own laboratories (28). Consequently, members have often been grudging in their support, which limits the projects that EMBL can undertake. EMBL's annual budget is approximately 45 million deutschmarks (about $26.5 million), 25 to 30 percent of which is paid by West Germany (68).

EMBL sponsors research in instrumentation, biocomputing, and gene mapping and sequencing as well as other areas of biology. EMBL's researchers have been active in technology development for mapping and sequencing and have produced prototypes of machines for automating some of the steps in DNA sequencing (see ch. 2).

EMBL also operates the major European database of nucleotide sequences, which works in cooperation with GenBank® to gather and disseminate sequence data. For EMBL to undertake a major human genome project would require a considerable increase in budget—unlikely under current circumstances—and sustained enthusiasm from its members [Newmark, see app. A]. Director-General Philipson is eager to promote collaboration on a genome sequencing project, which he believes will increase the need for a centralized European data-handling facility. In the 1986 director's report, Philipson encouraged the establishment of new support programs for a human genome project:

> If the American plan to launch a programme on the human genome materializes, the EMBL may be a natural collaborative partner in this project. It might, therefore, be worthwhile to plan for at least one new Programme in one of those fields to be initiated in Heidelberg at the end of the proposed Scientific Programme (1990). To facilitate recruitment and the launching of this Programme, plans should be available by 1990 but we do not foresee any cost during the next 4 years (36).

The European Science Foundation

Headquartered in Strasbourg, France, the ESF is subscribed to by 49 research councils and equivalent bodies from 18 European countries (33). It supports projects on a special funding basis from a small central fund; in the past, the ESF has not sponsored much research in biology, although recently it has supported some protein engineering work. One of the foundation's standing committees, the European Medical Research Council, enables the heads of national medical research bodies to meet once a year. The council has no budget, however, and little influence outside the ESF. At its 1987 meeting, the council decided not to attempt to coordinate European research on human genome mapping and sequencing [Newmark, see app. A].

The European Research Coordination Agency

A French-initiated response to the U.S. Strategic Defense Initiative, EUREKA was set up in 1985 to encourage development of advanced technologies in Western Europe. Participating in EUREKA are the 18 democracies of Western Europe: the 12 member states of the EEC (Belgium, Denmark, France, the Federal Republic of Germany, Greece, Ireland, Italy, Luxembourg, The Netherlands, Portugal, Spain, and the United Kingdom); the 5 member states of the European Free Trade Association (Austria, Finland, Norway, Sweden, and Switzerland); and Iceland.

EUREKA promotes industry-led technological collaboration among its members in several areas, including biotechnology and advanced information technology. It supplements EEC's efforts by funding research beyond the precompetitive stage. A EUREKA project must involve at least two industrial laboratories in two different European countries. Governments vary in their financial support of EUREKA projects: Some offer little more than token support and assistance in administering an international collaboration; others, such as France, pay up to 50 percent of a EUREKA project. Coordinated by a small secretariat in Brussels, EUREKA's performance has impressed many observers. Still, maintaining consistent funding is difficult, since most of the governments supporting EUREKA have not created procedures for funding the program (34). There are no EUREKA projects for human genome mapping and sequencing yet, but the program might be used to link French researchers to industrial partners in Europe, particularly in the development of sequencing technologies [Newmark, see app. A].

National Research Efforts in Europe

Denmark

The National Health Authority, the primary funding agency for biomedical research, supports some gene mapping studies, although there is at present no centralized effort. Other funds for gene mapping and sequencing come from general allotments to universities and research institutes, from the government, and from research councils, notably the Danish Research Council. Special projects can be funded by applying to the appropriate research council. The Institute of Medical Genetics of the University of Copenhagen is the most prominent Danish effort in the field. It has the longest tradition and the greatest interest in gene mapping; sequencing is not yet a major concern, although it may be in the future. A University of Copenhagen scientist is the editor of the international journal *Clinical Genetics*, which publishes mapping studies and similar research. There are several ongoing projects at the institute on various genetic diseases, but there is no concerted effort or government policy on mapping and sequencing (70).

One project of interest is a family pedigree project that has been underway for more than 10 years. Like the Venezuelan pedigree project (box 7-A), this is a collection of genetic material from families with many children; the collection contains "samples of red cells, serum, plasma, thrombocytes [parts of the blood that help in clotting], lymphocytes [cells important in the immune system], as well as skin biopsies" (59). Unlike the Venezuelan material, the genetic material in the Danish project was collected from apparently normal families; over the years it has been tested by classical genetic markers to help establish polymorphic regions for genes of different blood groups, enzyme types, and so on. Extensive RFLP mapping (see ch. 2) of the material has not been done because of limited resources, but negotiations are underway to contribute material to the Center for the Study of Human Polymorphism (CEPH), an international gene mapping center located in Paris, for further mapping. There is as yet no clear policy in Denmark on whether to sequence large portions of the family material, especially because resources are limited, but the research group is exploring the possibility of collaborative arrangements within Denmark, with other countries, and with the United States. The goal is to establish a Danish center for human gene mapping, LINK, starting with the family material that has already been gathered and expanding the collection, as well as drawing in researchers from other institutes. LINK is envisioned as a Scandinavian counterpart to the French CEPH effort (59).

The Danish Government has established 10 new biotechnological centers and allocated D.kr. 500 million (about $80 million) for their operating expenses over the next 5 years (6,59); 410 million will be used to establish new research centers at technical universities and private firms (24). The biggest center, at Aarhus, is already supporting some gene mapping research in collaboration with CEPH.

Federal Republic of Germany

The emergence of the environmentally oriented Green party in West Germany, combined with a general wariness about research with possible eugenic applications, has made molecular genetics research a sensitive political issue.[3] Nonetheless, research in molecular biology is well funded by federal, state, and private monies. There are four main sources of funds for basic research in molecular genetics. The Max Planck Society, which receives a substantial allotment from the federal government but is legally independent, supports the Max Planck Institutes, each of which is devoted to a particular area of research (72). The German Research Association (DFG) obtains approximately half of its funds from the federal government and half from state governments and supports research in the universities. The German Ministry for Research and Technology (BMFT) supports projects in universities as well as funding the Institute for Biotechnology Research and other research institutes. Individual states contribute to some science research through the universities. Another source of potential support for genome research is the prestigious Society for Biotechnological Research (GBF), a government-funded research center (78).

---

[3] One indication of this attitude is that a federally appointed commission of government and outside experts on genetic engineering recommended, in early 1987, that there be "tight limits drawn for analyses of human hereditary factors (genomic analysis) as well as for gene therapy" (2). The commission published an extensive report entitled *Chances and Risks of Genetic Engineering* after two years of study. An English translation of the foreword and recommendations of the report, entitled *Gene Technology: Opportunities and Risks* (16) has been made available by the EEC. The DFG criticized the recommendations of the commission in the case of genome analysis, arguing that the search for causes and cures for genetic defects is a scientific duty and serves the public interest (42).

At present, West Germany does not have a coordinated genome mapping or sequencing project. At a meeting in September 1987, representatives of the DFG decided not to endorse a concerted genome project, although the agency does support a research program targeting molecular methodology for studying the genome (52).

West Germans are strong supporters of international cooperation. They consistently contribute to EMBL, and several laboratories are carrying out research that could be extended at little expense and aligned with an international collaboration in genome research.

Biotechnology is being actively promoted by the federal and state governments in West Germany. The Federal Ministry of Research and Technology's Biotechnology Research Program, initiated in 1985, includes as an objective the promotion of "research and development projects in public life care, including health, nutrition, and environmental protection"; one of its high-priority research areas is a program of "genetic engineering with a focus on the investigation of gene structures, research on gene functions, and on controlling of genetic processes" (68). The ministry has also encouraged the establishment of research centers in which university and industry would participate and has set up seven "gene centers" to study areas including gene expression and differentiation and the correlation between gene structure and function. Human genome mapping and sequencing are not explicitly included in either the Biotechnology Research Program or the genetic research centers, but both support related research and could provide an institutional infrastructure and funding framework for genome research.

Finland

In January 1987, scientists at the Finnish Academy proposed a 5-year plan to improve biotechnology and molecular biology research, in order to promote industry and increase industrial capabilities. The proposal included a request for the equivalent of $37 million per year for research, training, and equipment (48). Finland has established several genetic engineering research centers and has plans for half-a-dozen more; the institute associated with the University of Helsinki is perhaps the best known.

Human genome mapping in Finland is being done by about 10 large and small individual research groups in medicine and science. They are primarily funded by government sources, namely, university budgets and the Academy of Finland, which is the main funding source other than universities. The University of Helsinki hosted the eighth international Human Gene Mapping Workshop (HGM 8) (5). Finland has no concerted effort nor any specific policies; as in most countries, however, sequencing efforts have focused on particular genes. Finnish groups are involved in collaborative projects with groups in other countries, notably the United States, and have contributed to and received materials from international databases and repositories.

France

Since 1981, the French Government has sought to make France a world power in science and technology by increasing both funding and political interest in research and development. The Government has encouraged collaboration between university and industry researchers, both within the country and with the rest of Europe (e.g., the EUREKA program).

The French Ministry of Research is directly or indirectly in charge of nearly all government-funded research. Most is carried out within universities, often in units set up by the research organizations, the largest of which is the National Center of Scientific Research (CNRS). The CNRS and the much smaller National Institute of Health and Medical Research (INSERM) are the only two government organizations that support research related to human genome mapping and sequencing. The Pasteur Institute in Paris, a semi-autonomous institute that receives half its funds from the government, carries out related research. None of these organizations has announced a firm plan for human genome mapping or sequencing, but each is considering what part it might play [Newmark, see app. A].

An important focus of genome studies in France is the CEPH (see box 7-B). Organized in 1983 by Jean Dausset to "hasten the mapping of the human genome by linkage analysis with DNA polymorphisms," CEPH is a privately funded center that collects and distributes genetic materials for use in mapping studies. It acts as an informal coordinator for approximately 40 investigators in Europe, North America, and Africa who use CEPH materials in exchange for reporting their data (25,26).

France has not initiated a coordinated genome project, but there is a strong undercurrent of opinion favoring a substantial program in human genome mapping and sequencing as long as it is not funded at the expense of other research. Genome researchers may try to work through EUREKA to involve other European companies with an interest in instrumentation or information technology. The French Government (usually through its Ministry of Industry) is prepared to provide 50 percent funding for EUREKA projects, and there are indications that it would consider CEPH's human genome work eligible for EUREKA funding [Newmark, see app. A].

Italy

Recent administrations have given priority to improving Italy's scientific performance in hopes of sparking a technology-led revitalization of the country's ailing economy. Considerable extra funds for technology-related research have been made available in the past few years, with biotechnology as one focus. The Italian Government announced in April 1987 that it would allocate 209 billion lire (approximately $156 million) over a 5-year period for a national biotechnology project involving both public research centers and industry (64); the following month Italy's National Research Council (CNR) announced a special research project in biotechnology for which it will spend 84 billion lire (about $63 million) over the 5-year period (51).

In May 1987, the CNR announced its decision to initiate a project devoted to human genome sequencing, to be run as a cooperative effort of all CNR institutes and laboratories working in biology (22). Nobel laureate Renato Dulbecco is coordinating the project, in which CNR has started investing 20 billion lire (about $15 million) and 75 to 100 person-years (51). A 2-year pilot project with a budget of $1 million per year will be undertaken first, to determine whether a large-scale project will be funded at around $10 million a year. (These sums are to cover only specific materials, machines, travel, meetings, and so on—not salaries and general overhead—since only the existing number of personnel will be involved.)

A key question in the pilot project is whether it is possible to isolate a single chromosome without damaging it so much that sequencing would be impossible. The ability to separate the chromosomes would offer a shortcut to sequencing, and researchers could begin sequencing with one of the smaller chromosomes (but one with genes of particular interest), probably chromosome 21, 22, or Y (73). Otherwise, researchers will consider continuing the project using conventional techniques. Research institutes and laboratories in Rome, Naples, Pavia, and Milan will participate in the project. Databases and information retrieval will be managed by research units in Rome, Turin, Milan, and Bari, with the aim of making the national databases compatible with and complementary to existing international ones (57,73).

The pilot human genome project is still exploratory, so no attempt is being made yet to coordinate work with researchers outside Italy. Project scientists anticipate that the final project would be complementary to, if not an integral part of, any international project that arises [Newmark, see app. A]. In the meantime, Italian scientists are enthusiastic about Italy's role in genome mapping; "there is good reason to believe that, for once, this country will perhaps succeed in reaching the starting line ahead of other countries" (73). Italian scientists are not the only ones interested in chromosome 21, however; it is a popular target for research because it contains genes for Alzheimer's disease and for Down's syndrome, and it is likely to be an early focus of U.S. efforts.

Box 7-B.—The Center for the Study of Human Polymorphism (CEPH): An International Gene Mapping Center

The Centre d'Etude du Polymorphisme Humain (CEPH) has become an important focus of international scientific cooperation in the drive to map the human genome. CEPH is a private research foundation established in 1983 by French Nobel laureate Jean Dausset with the bequest of an anonymous donor. Its aim is to "hasten the mapping of the human genome by linkage analysis with DNA polymorphisms."

The basic premise behind CEPH's activities is that a genetic linkage map (see ch. 2) will be more easily constructed if researchers study genetic material from a common group of families—a reference panel. The most useful family pedigrees consist of four living grandparents with many children and grandchildren so that the inheritance of DNA can be traced through three generations. CEPH maintains DNA from a panel of 40 families, each with 5 to 15 children; in most cases, all grandparents are living. The DNA from 29 of the 40 families in the CEPH collection was contributed by Ray White and his collaborators from the Howard Hughes Medical Institute (HHMI) in Utah. Dausset also solicited family materials gathered by other researchers in the United States and Europe, including some material from normal families identified in the Venezuelan pedigree project. In contrast to that project (see box 7-A), in which researchers collected material from families with Huntington's disease in order to trace the gene responsible, CEPH maintains material from families with no known genetic diseases. The markers mapped to chromosomal locations in normal CEPH families can then be used to accelerate the search for disease genes in other families.

CEPH coordinates an international collaboration of researchers from laboratories in Europe, North America, and Africa. In order to obtain material from CEPH, collaborating investigators must first possess DNA probes that detect genetic markers, generally RFLPs. They must agree to use the probes to test the entire panel of 40 families and to provide CEPH with all of their data. There are no enforcement mechanisms, but so far researchers have cooperated.

Dausset's work is supplemented by the efforts of Jean-Marc Lalouel, a mathematical geneticist at HHMI in Utah who has designed a variety of computer programs to record and analyze the data contributed by CEPH investigators. Lalouel and his collaborators have written programs that analyze genetic linkages and automatically sketch out gene maps from the results. These programs are sent out on disk with the CEPH DNA samples. Researchers can record and analyze their data using the programs on the disk, then send the disk back to CEPH for inclusion in a central database. HHMI supports a database station at CEPH that will be linked to its Utah station and may soon include interactions with other databases as well.

An important factor in CEPH's success at fostering cooperative research is the two-tiered database it maintains. One database, available only to collaborators, contains all data that investigators produce. At the end of a year's time or when the results have been published, whichever comes first, data from the collaborative database is moved into a public database, where it is accessible to any qualified researcher. This system of having both a private and a public database ensures the timely sharing of information while affording investigators some proprietary protection for their results. The fact that the collaboration requires sharing of data—but not the actual probes, which could prove to be patentable—reduces potential competitive tensions.

SOURCES:

H.M. Cann, "Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative Mapping of the Human Genome," paper submitted in preparation for the Oct. 16-17, 1986 meeting of the Advisory Committee to the Director, NIH.

H.M. Cann, CEPH, personal communication, December 1987.

Centre d'Etude du Polymorphisme Humain, unpublished report, 1986.

J. Dausset, H. Cann, and D. Cohen, "Centre d'Etude du Polymorphisme Humain (CEPH): Collaborative Mapping of the Human Genome," unpublished manuscript, April 1987.

J.L. Marx, "Putting the Human Genome on the Map," *Science* 229:150-151, 1985.

M. Pines, *Mapping the Human Genome* (Bethesda, MD: Howard Hughes Medical Institute, 1987).

R. White, M. Leppert, D.T. Bishop, et al., "Construction of Linkage Maps with DNA Markers for Human Chromosomes," *Nature* 313:101-105, 1985.

Industry is not playing a role in the pilot project, since few Italian companies have the technological interest or capability. But scientists involved in the research believe that "the automation required for the project will act as a major incentive for industry" and hope that industry would help finance the final project (73). At least one Italian pharmaceutical company has expressed a willingness to participate and contribute.

The United Kingdom

The United Kingdom has a strong research tradition in molecular biology and genetics, and it has done pioneering work in the mapping of nonhuman genomes and in the development of sequencing techniques. The United Kingdom has consistently ranked second to the United States in the number of articles on human gene mapping and sequencing published annually in international journals (see figure 7-1, table 7-1). The United Kingdom also ranks high in the development of physical mapping techniques and of automated technologies for DNA manipulation and analysis. Thus the United Kingdom is well placed intellectually, if not financially, to contribute significantly to mapping the human genome.

Basic biomedical research is funded mostly by the government through the Department of Education and Science, although both the Department of Health and Social Security and the Department of Trade and Industry have funds available for contract research. The Department of Education and Science distributes research monies through universities and through five research councils. The research councils provide support for scientific programs carried out in universities; some councils also support research within their own institutes. Biotechnology is an area of overlap for the Science and Engineering Research Council (SERC) and the Medical Research Council (MRC), the two councils whose areas of interest are most closely related to human genome research. The science and engineering council supports basic biological research outside the medical field, although it has supported some work on automated DNA sequencing through a biotechnology directorate established to link academic research to industrial needs. The MRC is undoubtedly the leading supporter of mapping and sequencing research. Its total expenditure for genome-related research for the 1985-1986 fiscal year, both direct and indirect, was approximately £4.2 million ($7.4 million) [Newmark, see app. A] (88).

The MRC is similar to the NIH in supporting high-quality, investigator-initiated proposals, although the council also establishes targeted programs in particular areas. It has a longstanding commitment to molecular biology and has the power to set up new units devoted to particular areas of research when a suitable director and sufficient funds are available. Although the MRC supports a good deal of relevant research and its various units and grant holders have the expertise and instrumentation necessary for the study of genetic disease, the MRC does not now plan a targeted program of research on human genome mapping or sequencing. At a 1987 meeting, however, the MRC did endorse the plan of an employee, well-known scientist Sydney Brenner, to map the human genome (largely with private funds) as long as the research proceeded at no extra cost to the research unit Brenner directs (66). At Brenner's request, the MRC has also agreed to set up a committee that will consider questions such as who owns the clones produced in mapping efforts and how best to provide public access to them.[4]

Brenner's project will be financed in part by a £300,000 (about $525,000) prize award he received from the Louis Jeantet Foundation; the MRC and other sources will provide another £200,000 to £250,000 (about $350,000 to $440,000) per year (56). The project will build on a mapping technique developed by Alan Coulson, John Sulston, and co-workers in the MRC research unit at Cambridge. They compiled a genetic linkage map of the nematode *Caenorhabditis elegans* genome, the smallest genome known for any multicellular creature (it is estimated to be 80 million base pairs, compared with approximately 3 billion base pairs for the human genome—see ch. 2). Brenner expects that perhaps half of the genome could be mapped by a few people within 5 years. The project will include research on data-handling methods and parallel processors, since the mapping techniques require sophisticated computing capabilities.

---

[4] "It has been agreed [by the MRC] that the human genome work should constitute a separate project to be carried out as an extension of the work of the [Molecular Genetics] Unit [in Cambridge]. It was also considered that the longer term future of this work could not be tied to the finite tenure of a personal Unit. The project might evolve into a reference laboratory with a major service component and would then need a different funding structure. A central aim would be to ensure that the collection of clones and information remained in the public domain. It was therefore agreed that an Advisory Board be established to consider these and other policy matters" (66).

The Imperial Cancer Research Fund (ICRF), a charitable organization financed solely by donations, has recently recruited scientists to work on the development of a different technique for human genome mapping, as well as related software and instrumentation [Newmark, see app. A]. The MRC and ICRF plan to explore the possibility of collaboration in areas of common interest.

Other efforts in the United Kingdom include technology development in automated systems for genome sequencing at the University of Manchester Institute of Science and Technology (UMIST) (1) and biocomputing research at the University of Edinburgh. The Edinburgh Biocomputing Research Unit has considerable experience in database searching and related problems and is undertaking a variety of studies into the informatics needed for analysis of map and sequence data (15).

The United Kingdom contributes to international research efforts such as EMBL, to which the MRC provided £2.72 million (about $4.7 million) in 1987. The MRC maintains a level contribution to EMBL in real terms, after supporting some growth of the organization in 1982, when the new director was appointed (66).

OTHER INTERNATIONAL EFFORTS

Australia

The largest research institution in Australia is the Commonwealth Scientific and Industrial Research Organization (CSIRO), which is conducting pertinent research through its Division of Molecular Biology. Biomedical research is primarily the province of the National Health and Medical Research Council, which at present funds a number of researchers working on gene mapping and sequencing. The Department of Human Genetics and the Medical Molecular Biology Unit at the Australian National University in Canberra are sites of some relevant research activity. In particular, chromosomes 6 and 9 are the foci of investigation because several genes have been localized to them (43,82). Researchers at the Cytogenetics Unit Department of the Adelaide Children's Hospital in North Adelaide are constructing maps of chromosome 16 and part of the X chromosome. They have collaborated with scientists from the U.S. Department of Energy's Lawrence Livermore and Los Alamos National Laboratories.

The Department of Industry, Technology and Commerce administers a system of research grants under its National Biotechnology Program, with priority areas including genetic engineering and cell manipulation and culture, which could provide support for genome research.

Canada

Canada does not yet have a national policy on genome sequencing. The National Research Council (NRC) is considering the creation of a task force to address this subject within its laboratories. A national network of biotechnology laboratories supported by the council has been set up, including the Biotechnology Research Institute in Montreal, the Plant Biotechnology Research Institute in Saskatoon, and the Division of Biological Sciences in Ottawa, which focuses on protein engineering.

In addition to the expertise that the government research institutes might lend to genome research, Canada has 15 to 25 university laboratories with the necessary skills and equipment to participate in a human genome project. To date, however, there has been little effort to coordinate the activities of these various groups. Canadian scientists and government officials are paying close attention to international developments in human genome sequencing and are hopeful that opportunities for international collaboration will develop (67).

Latin America

Relatively few laboratories are involved in human genome research; of those that are, the primary interest is generally mapping genes for diseases of particular national significance. As one observer pointed out, "Brazil has its share of good scientists, but they are hampered by lack of funding and difficulties importing equipment and materials"; presumably the same holds true in other Latin countries (13).

Many Latin American countries realize the commercial potential of biotechnology; Brazil and Argentina, among others, have initiated programs to encourage biotechnology research and development. Argentina has a biotechnology program under the aegis of its Secretariat of Science and Technology (60), and Brazil has a Biotechnology Secretariat in its Ministry of Science and Technology (13). Scattered throughout Latin America are individual laboratories doing relevant research.

In Mexico, "scientists are pushing the Mexican government to consider the development of genetic research a priority. They don't want to fall behind on this kind of research, because the pathology index in the Mexican population is approaching that of developed countries. With epidemics and infections decreasing, greater attention can be paid to genetic problems" (69). Like Brazil, however, Mexico has a low research budget (less than 0.6 percent of the gross national product is spent on research) and can neither afford sophisticated equipment nor train enough scientists; both countries are interested in international cooperation. The Organization of American States reports that its Department of Scientific and Technological Affairs, which runs a Regional Program for the Development of Science and Technology in Latin America and the Caribbean, includes projects in plant and animal genetics but none in human genetics (65).

South Africa

Gene mapping and sequencing research is supported by the Medical Research Council (MRC), the Council for Scientific and Industrial Research (CSIR), and the National Cancer Association. None has initiated a formal or coordinated attempt to map or sequence the human genome, but there are a number of laboratories at work in the field of human genetics (30). Several researchers are active in the CEPH collaboration, screening the CEPH family materials and contributing their results. Researchers are examining genes for Huntington's disease, cystic fibrosis, and neurofibromatosis in collaboration with laboratories in the United States and the United Kingdom (10). Research is also underway on several genes of particular interest in the region—those for Oudtshoorn skin disease and familial hypercholesterolemia (conditions prevalent in Afrikaners) and albinism, which is common in the Bantu population (50).

The Union of Soviet Socialist Republics and Eastern Europe

Although the Soviet Union has not been a major contributor to mapping and sequencing studies published in international journals, it has published some research on bacterial genomes (74) and the barley genome (3). Soviet scientists are also working on computational methods for analyzing DNA sequences (7). The Central Institute for Molecular Biology in East Berlin has undertaken a variety of studies in gene mapping and sequencing and has collaborated with researchers in the United Kingdom (45). Bibliometric analyses (see figure 7-1 and app. C) show that the Soviet Union and Eastern European countries have not published a significant number of research articles on human gene mapping and sequencing. These figures tend to select items from international journals, however, so internal publications are not as thoroughly catalogued and accounted for.

INTERNATIONAL COLLABORATION AND COOPERATION

The large size and humane mission of human genome projects make them ideal candidates for international collaboration. International databases have already been established and are being jointly maintained, which indicates some willingness to cooperate on gene mapping efforts, but it remains to be seen how far that cooperation will extend. The potential for commercial payoffs raises difficult questions but does not preclude successful collaboration as long as prior agreement on allocation of benefits is reached (32,49). The following sections recount some precedents for collaboration and cooperation in international science projects and the role the United States has played in them. Organizational options available for international human genome projects are examined, and some collaborative efforts already underway are described. The following chapter outlines the questions of international technology transfer that will undoubtedly arise in any coordinated international effort.

Precedents for International Scientific Programs

The biological sciences have been organized into international projects far less often than other sciences, but collaborations in the physical and space sciences can provide useful organizational insights. The International Geophysical Year, box 7-C, is an example.

Since the 1940s, research in particle and high-energy physics has relied on complex and expensive equipment—notably, the particle accelerator—that is beyond the ability of any individual investigator, or even any one institution, to construct and maintain. Consequently, a number of large, specialized laboratories have emerged nationally and internationally. In the United States, centralized facilities evolved into a network of national laboratories, now operated by DOE. These laboratories house cyclotrons, synchrotrons, and other advanced instruments and undertake research in a broad range of areas, cooperating in limited ways with researchers from abroad.

The European Center for Nuclear Research (CERN) was established in 1954 to advance knowledge in the field of particle physics. It is operated by 14 European nations and has provided a framework for collaboration in instrumentation. Its governing council consists of one technical advisor and one administrative advisor from each member nation. Participants contribute to CERN based on their gross national products, although no nation can contribute more than one-quarter of CERN's annual operating budget. CERN has enabled European nations to conduct research beyond the capabilities of any single member nation and has been widely recognized for its success in the advancement of particle physics. It has restricted its efforts to basic research, however, and so has avoided the complications that arise in collaborative work on applied research (80).

The enormity of the endeavor to explore and study space spawned proportionately large agencies to manage the research. The founding legislation of the United States' National Aeronautics and Space Administration (NASA) included international cooperation as a major theme, and NASA has carried out that mandate by negotiating and implementing hundreds of cooperative projects. Some NASA projects have established formal joint working groups on a bilateral basis with other national agencies. These groups meet several times a year to "discuss present and future projects of mutual interest, and to exchange information on scientific and management issues of concern" (61).

One of NASA's major partners has been the European Space Agency (ESA), a collaboration of 13 European nations. The Hubble Space Telescope is an example of collaboration between the two agencies. In 1977, officials from NASA and ESA drew up an agreement to work together on the project, citing specific contributions and responsibilities (37). An article on data rights directed that scientific data from the telescope be reserved for analysis for one year, then turned over to public data centers. Results were to be made available to the scientific community through publication as soon as possible and appropriate. No specific provisions were made for patenting products or processes developed in the course of the project.

Box 7-C.—The International Geophysical Year

The International Geophysical Year (IGY) was originally conceived as the third in a series of international polar years—earlier cooperative investigations into the phenomena of the Arctic and Antarctic took place in 1882-1883 and 1932-1933—but the scope was expanded to include the study of all aspects of the physical environment. Sydney Chapman, one of the organizers, described the enormous undertaking as it finally evolved:

> The main aim is to learn more about the fluid envelope of our planet—the atmosphere and oceans—over all the earth and at all heights and depths. The atmosphere, especially at its upper levels, is much affected by disturbances on the Sun; hence this also will be observed more closely and continuously than hitherto. Weather, the ionosphere, the earth's magnetism, the polar lights, cosmic rays, glaciers all over the world, the size and form of the earth, natural and man-made radioactivity in the air and the seas, earthquake waves in remote places, will be among the subjects studied. These researches demand widespread *simultaneous* observation.

To accomplish this, teams of scientists from 67 nations—60,000 in all—observed, measured, and recorded data in meteorology, geomagnetism, auroras and airglow, the ionosphere, solar activity, cosmic rays, oceanography, glaciology, gravity measurements, and other disciplines over a period of 18 months in 1957 and 1958.

The effort was coordinated by the Special Committee of the IGY (CSAGI) under the auspices of the International Council of Scientific Unions. Planning committees were appointed to organize research programs in 14 different disciplines. Participating nations generally had their own planning commissions or advisory boards as well.

An essential feature of IGY was the operation of world data centers. Participants agreed to send all of their data to three major centers, in the United States, the U.S.S.R., and Western Europe. Organizations or investigators from any country could obtain copies of the deposited materials free of charge (other than the price of reproduction and transmission). In addition, the data were summarized and presented in more than 30 volumes in the *Annals of the International Geophysical Year*, an information resource that provided the raw material for subsequent research in geology, meteorology, oceanography, and other fields.

SOURCES:

S. Chapman, *Annals of the International Geophysical Year*, forward quoted in H. Newell, *Beyond the Atmosphere: Early Years of Space Science* (Washington, DC: NASA, 1980).
S. Chapman, *IGY: Year of Discovery* (Ann Arbor, MI: University of Michigan Press, 1959).
J.M. England, *A Patron for Pure Science: The National Science Foundation's Formative Years* (Washington, DC: National Science Foundation, 1982), pp. 297-304.
H.E. Newell, *Beyond the Atmosphere: Early Years of Space Science* (Washington, DC: NASA, 1980).
W. Sullivan, *Assault on the Unknown: The International Geophysical Year* (New York, NY: McGraw-Hill, 1961).

NASA's operating principles for international collaboration are a useful starting point for drawing up collaborative agreements.[5] One key difference, however, between human genome projects and most space research is the commercial potential: "Astronomical data have no commercial value" (71). The gap between research in molecular genetics and the market has narrowed rapidly in recent years, making the boundary between basic and applied or development-oriented research nearly impossible to draw. Consequently, agreements similar to those negotiated by NASA and ESA regarding data rights and publication of results could prove insufficient for human genome projects. A second difference is that the instrumentation required for human genome projects is neither as large nor as expensive as that used in particle physics and space research.

In spite of a stated desire for international cooperation, the United States has generally acted as the primary partner in large science projects, defining them and then inviting other nations to join in, rather than planning, funding, and implementing projects jointly (54). In the present era of constrained funding, however, the United States may not always be able to carry out major research projects on its own.

---

[5] NASA has never formally encoded its mechanisms for international collaboration, but it has developed an informal set of guidelines:

- Cooperation is on a project-by-project basis, not on a program or other open-ended agreement.
- Each project must be of mutual interest and have clear scientific value.
- Technical agreement is necessary before political commitment.
- Each side bears full financial responsibility for its share of the project.
- Each side must have the technical and managerial capabilities to carry out its share of the project; NASA does not provide substantial technical assistance to its partners, and little or no U.S. technology is transferred.
- Scientific results are made public (55).

Collaborative projects can offer significant savings for participating countries by splitting the financial burden (although some observers have pointed out that the costs of negotiating and the loss of jobs if a project is located outside the United States may reduce the savings). Collaboration creates a paradox, however: On the one hand, it might reduce the cost for each member, making the project more feasible; on the other, it might reduce each nation's potential economic gain from the project. The world economic situation has led to an increasing desire for scientific research to produce commercially valuable products, thereby fostering a protective, nationalistic attitude toward research (see box 7-D).

Options for International Organization of Genome Research

A decision to pursue human genome projects on the international level, emphasizing cooperation and participation, will entail considerable organizational effort. It will have the same organizational goals as a domestic effort: to eliminate redundancy in research and to expedite the spread of scientific and commercial knowledge of the ge-

Box 7-D.—Views on International Cooperation and Collaboration in Genome Research

"Too many promising international research collaborations, from AIDS research to the sequencing of the human genome, languish for lack of a workable framework for tangible and short-term research. . . . The U.S. Department of Energy and the Japanese Science and Technology Agency have an interest in organizing and supporting the [genome] project; each seems sensibly to have decided that two independent projects would be a waste of resources and a source of confusion, but [they] differ sufficiently in their objectives as to impede agreement between themselves, let alone with others." Editorial in *Nature* 328:187, 1987.

"There's a task to be done here, and we need to get on with the task. If we try to take into account every country's interest and concerns, we can only serve to delay it." J. McConnell, Johnson & Johnson, Science Writers' Workshop, Brookhaven National Laboratories, Upton, NY, Sept. 14, 1987.

"An international DNA analysis center or centers equipped with super sequencing systems which are connected to a worldwide data-network should be developed." A. Wada, "Many Small-Scale or a Few Large-Scale DNA Sequencers?" unpublished report, Japan, 1987.

"It is highly desirable that the U.S. continue to be the leader of the [genome mapping and sequencing] effort, but it must be consciously and effectively run as an international quest for knowledge having universal importance. No single purse nor administrative center, in either the U.S. or the world, can or should be created to fund or attempt to direct the task." D. Fredrickson, National Institutes of Health, personal communication, December 1987.

"There is . . . a growing awareness in Europe that the first megaproject in biology is shortly being launched. Europe ought to participate in it alongside the USA and Japan to ensure access to the information and all that it implies for medical and biological science, as well as the technological spinoffs that will surely arise. . . . There is now an opportunity to ensure that the project involves international collaboration from its outset which should not be missed." L. Philipson and J. Tooze, "The Human Genome Project," *Biofutur*, June 1987, pp. 94-101.

"If they wished, either Western Europe or Japan could by themselves take on this project and it must be assumed that they will initiate their own efforts. So a new international body should soon be formed to ensure that collaboration, not competition, marks the relationship between these efforts in various parts of the world. In a real sense, the exact sequence of the human genome will be a resource that should belong to all mankind. So it is a perfect project for us to pool our talents, as opposed to increasing still further the competitive tensions between the major nations of the world." J.D. Watson, director's report for Cold Spring Harbor Laboratories, in press.

"The principle of 'mutual self-interest' . . . lies at the heart of successful cooperation." D. Dickson and C. Norman, "Science and Mutual Self-Interest," *Science* 237:1102, 1987.

"If a sequencing factory can be built, Wada emphasizes that it would not be 'Japan Incorporated' against the rest of the world. He wants an international centre that would be open to scientists of all nationalities and intended for the benefit of all mankind." D. Swinbanks, "Human Genome: No Consensus on Sequence," *Nature* 322:397, 1986.

"This project is so vast that it necessarily requires international cooperation. Since there are 3 billion bases to be sequenced, the project will not create problems of competition." P. Vezzoni, Consiglio Nazionale di Richerche, Milan, quoted in A. Sommariva, "And Italy Will Study Chromosome 22," *Italia Oggi* (Milan), May 22, 1987, p. 36.

"There's considerable interest in the commercial spinoffs, and I expect each country would want to keep those. I would hate to see U.S. tax dollars used to kill yet another U.S. industry." J. McConnell, Science Writers' Workshop, 1987.

"On the one hand, the climate for international collaboration in science . . . is warmer than ever. In virtually every major field, U.S. scientists can point to significant work being done in Europe, the Soviet Union, Japan, Canada, or Israel that needs to be read closely, argued about, and replicated as much as does work done in the United States. On the other hand, the new era is chillier, for governments and businesses here and abroad will continue to try to squeeze economic value out of every bit of science to win the international high-tech sweepstakes." D. Shapley and R. Roy, *Lost at the Frontier: U.S. Science and Technology Policy Adrift* (Philadelphia, PA: ISI Press, 1985), p. 116.

"The creation of a sequence database is the major goal of the project, whether it is done nationally or internationally or privately. . . . I don't think an international project as an organized scheme will emerge. . . . I expect a set of private ones will emerge, with some level of cooperation." W. Gilbert, Harvard University, Science Writers' Workshop, Brookhaven National Laboratories, Upton, NY, Sept. 15, 1987.

"I am convinced that an international advisory body must be formed to oversee the data bases. . . . International cooperation [is] as important as interagency coordination in the U.S.A. But I do not think that a special institution would be useful at the national and at the international level." A. Lafontaine, Office of the Secretary General, Brussels, Belgium, personal communication, June 1987.

"There is a strong belief here that practical collaborations on actual, well-defined projects are very helpful, and probably more meaningful than large-scale collaboration between governments. Cell banks, gene banks, and databases are very important in this regard." A. de la Chapelle, University of Helsinki, personal communication, August 1987.

"International cooperation is not something that should be imposed by government agencies. . . . Real cooperation comes from individual scientists communicating with each other." C. DeLisi, Mount Sinai School of Medicine, Science Writers' Workshop, Brookhaven National Laboratories, Upton, NY, Sept. 15, 1987.

"I'm just concerned that if we focus on trying to set up an international effort, we will delay decisions of the United States in proceeding with this. I'd like to see a willingness to cooperate at the international level, but setting U.S. national priorities." G. Cahill, Howard Hughes Medical Institute, comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.

"[T]he United States does not and cannot expect to monopolize information and innovation in this field. Moreover, the initiation of a human genome project in the United States will probably not deter work in other countries, but rather will stimulate it. Given this assumption, the importance of past traditions, and the magnitude of the task of mapping and sequencing the entire human genome, every effort should be made to enhance the existing contacts between the United States laboratories and those overseas, so as to speed the work. Indeed, we believe it will become necessary to have some major organized mechanism for international cooperation. In particular, its objective would be to collate data and ensure rapid accessibility to it, as well as to distribute materials, such as cloned DNA fragments." National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988), p. 85.

nome. Just as the issues in domestic organization revolve around distribution of authority and tasks among interested government agencies and private firms (described in ch. 5), the issues in international organization involve coordination of interested sovereign nations.

An international organization could be either passive or active. A passive organization would serve primarily as a clearinghouse of research information among participating nations. This task would require the formulation and oversight of standard nomenclature and the translation of research reports. The organization would need to keep track of research in progress and any technological innovations reported by individual laboratories, and it might be intimately associated with databases such as GenBank® and the EMBL data bank and with collaborative organizations such as CEPH. Although participation in this type of passive organization would have to be voluntary, all academic researchers would stand to benefit from the free flow of information. The proprietary interests of commercial researchers might limit their participation, but collaborative arrangements could be made (12,49). The success of a passive international organization depends primarily on the good will of the participants.

An active international organization along the lines of the interagency task force described in chapter 6 could plan and distribute genome research among participating countries. **There are at least three ways in which the tasks of an international genome project may be distributed: 1) by physical units**, such as chromosomes or genes, in which each country would analyze one unit or a group of units; **2) by project aspect**, such as sequencing, informatics, or cloning, in which each country would focus on one aspect; and **3) by geography**, in which each country or group of countries with similar resources would establish a genome center.

Distribution by physical units would require each participating nation to possess the entire spectrum of technical specialties associated with the project—mapping, sequencing, data management, and so on. This requirement would probably limit involvement to those nations that are already scientifically advanced, regardless of any interest among nations attempting to develop biotechnical capabilities. The requirement could, however, spur developing nations to acquire technologies, and it might provide an economic incentive for commercial firms to assist in the start-up efforts. Assignment by chromosome would most likely cause intense politicking among the top nations for the most "interesting" chromosomes. Certain countries or regions might be more interested in chromosomes known to contain genes that affect a large portion of their populations. Such a method of assignment would also identify a specific nation with a specific achievement, effectively placing flags on the map of the genome. The realization of this would inject an element of competition for national prestige into the context of an international science project. In effect, the cooperative partners would be establishing the arenas and ground rules for competition.

An international project divided by project aspect would require participating countries to adopt a specialty, which would accelerate development and commercial profit in that field but could preclude achievement in related fields. Japan, for example, might contribute a large share of DNA sequencing because of its interest in automating sequencing technologies. The component tasks of a genome project are not equivalent nor easily evaluated in terms of necessary resources, so distributing them may prove difficult. Further, some aspects of the project are more visible and economically valuable than others. To map or sequence an important gene is noteworthy and profitable; to create a database is to provide a common good but to receive little of value in return. An international division of labor is an attractive idea, but only clearly defined special talents among the nations would justify it.

The third possible distribution of international efforts is geographical—several genome centers could be established and supported by a nation or group of nations. The vocation of these centers might become a point of debate, however: Should each cover the full spectrum of genome technology, or should they specialize?[6] If each center attempted to cover all technologies, a division of labor might evolve based on specialized innovation. This might keep the centers complementary and competitive, but not necessarily cooperative. Establishing specialized centers would predetermine each center's scientific and economic success. Focusing all of them on a single aspect, for example sequencing, would siphon funds and attention from the other aspects. A center arrangement involving only countries with state-of-the-art research capabilities might lock out interested countries just beginning to develop biotechnology capabilities, unless the centers were amenable to taking on minor partners. Few scientists other than the two who have proposed the sequencing center idea seem to be enthusiastic about the prospect of establishing large centralized institutions (see box 7-D).

---

[6] The idea of setting up large centers has been promoted by American scientist and entrepreneur Walter Gilbert (44) and by Japan's Akiyoshi Wada (84,85,87). Both have referred specifically to sequencing rather than to genome research in general.

If an international project is to be pursued, issues of participation and underlying motivations should be recognized clearly and early. Without specific guidelines for initial and future participation, any organization is likely to become entrenched and inaccessible to latecomers. If the motivation for an international distribution of effort is purely economic, then participation might be restricted to nations already able to demonstrate their ability to contribute. Should an international effort be tied to political goals such as assisting the growth of biological research and biotechnology in the developing world, then widespread participation and an organization capable of coordinating both advanced and developing countries would be necessary. If political motives are acknowledged, then the international organization might seek to encourage the association of national goals and priorities with genome research. Political motivations are probably inherent in international projects, but they could be used to elicit widespread participation and continuing commitment. By using enticements such as distribution of physical units of the genome by political units of the participants, it may be possible to guide nationalistic forces into a workable international effort.

An important factor in any international collaborative or cooperative agreement will be the participants' domestic organization of human genome projects. The agencies involved speak with many voices, depending on their respective missions. Formal collaboration would be difficult to negotiate without some domestic coordination (see chs. 5 and 6) to harmonize goals. Otherwise, less formal cooperative arrangements will probably prevail.

Even if there emerges no formal international organization that can satisfy national and proprietary goals, the United States could establish an international advisory board to solicit suggestions and recommendations from the international scientific community regarding human genome projects. Domestic advisory boards could include nonvoting members from Europe and Japan. An international advisory committee for database oversight already exists; it has two members from the United States, two from Japan, and several from Western Europe (14). Members of the committee issue recommendations that, although not binding, help coordinate the various national efforts.

Existing Collaborative Frameworks

Lack of an international organizational structure does not preclude informal collaboration or cooperation. Scientific laboratories exchange views, visits, and materials as a matter of daily practice; many scientists prefer informal networking to prescribed arrangements and institutions (see box 7-E). Policymakers in Europe are finding that increasing support for laboratory networks, rather than establishing centers, can be an effective way to conduct research on a limited budget. Many of the scientists involved in human genome research host visiting foreign scientists and graduate students regularly.

The United States already finances international collaboration in biomedical research to a certain extent through the normal funding mechanisms of the National Institutes of Health, which may award grants to U.S. investigators "whose work involves substantial collaboration with foreign institutions" (63). Researchers affiliated with foreign institutions are eligible for grants and contracts; in fiscal year 1984, NIH spent $35 million on foreign grants, roughly half the budget allotted to international activities. NIH also gives grants for foreign or international conferences and for international research fellowships.

Box 7-E.—Large Centers v. Networking

The development of international sequencing centers draws enthusiastic response from some quarters and skepticism from others. Proponents such as Walter Gilbert and Akiyoshi Wada advocate the creation of several international centers containing advanced sequencing equipment as the most efficient way to sequence the genome, if not to map it. Critics contend that establishing large central institutions reduces the innovation spawned by small research laboratories doing investigator-initiated projects. Other critics, including many industrialists, argue against "naive internationalism," stating that the task at hand should be done posthaste, without lengthy delays while international negotiations decide on the division of labor, responsibilities, and benefits.

One solution that could satisfy critics of both stripes is networking—strengthening the links between existing laboratories—rather than starting up new research centers. Networking has recently gained popularity in the European community; indeed, Dickson has written that "the top political priority given to the idea that governments should focus their efforts on linking together scientists in existing laboratories—rather than on creating major centers or research facilities—has become perhaps the most important shift in European-level science policy in the 1980s."

Various research programs supported by such organizations as the European Science Foundation and the European Economic Community (EEC) have adopted networking strategies in lieu of costlier and more contentious decisions to set up central collaborative facilities. The ESF has supported laboratory networks for research in areas including polar science and individual psychological development. Particularly relevant for genome research is a network on the molecular neurobiology of mental illness, in which scientists are hunting for pedigrees of families with psychiatric problems in order to locate informative genetic polymorphisms for linkage analysis studies (see ch. 2 and box 7-A). The EEC supports research under the Stimulation Program, providing money to allow scientists from different countries working on the same project to meet, perform joint experiments, and so on. One successful project that the Stimulation Program funded, according to Dickson, was "a research program into the development of new high-field magnets, which now links together scientists working in 58 research institutions in the 12 member states of the EEC. The EEC's Biotechnology Action Program, which encourages a transnational approach to the research it sponsors, has developed a similar networking approach—European Laboratories Without Walls (ELWWs). ELWWs link individual researchers from laboratories in different institutions (preferably in more than one country) together for multidisciplinary but focused, precompetitive research projects. The ELWW program emphasizes rapid, open flow of information and material between participants and incorporates joint planning and evaluation of the scheduled experiments.

Perhaps because European laboratories have traditionally been poor at communicating beyond their national borders—European scientists are more likely to collaborate or cooperate with American scientists than with other Europeans—the networking strategy has met with increasing enthusiasm and has fostered notable successes. Whether the strategy would work to link Europe, Japan, and the United States is not certain. Even within Europe there are potential problems. Networking could lead to the support of elite research groups and exclude those from poorer countries that do not yet have the facilities to be desirable research partners. For projects with potential commercial value, proprietary rights and the open exchange of information can become troublesome issues. Dickson reports that some policymakers argue that "the relative absence of centralized strategic thinking could turn out to be a major weakness." Despite these caveats, networking is a model for international organization that could reduce the anxieties accompanying the planning and implementation of international cooperative or collaborative projects.

SOURCES:

D. Dickson, "Networking: Better Than Creating New Centers?" *Science* 237:1106-1107, 1987.

J. Heilbron and D. Kevles, see app. A.

J. Maddox, "New European Collaborations," *Nature* 330:417, 1987.

J. McConnell, Johnson & Johnson, Science Writers' Workshop, Brookhaven National Laboratories, Upton, NY, Sept. 14, 1987.

R. van der Meer, E. Magnien, and D. de Nettancourt, "European Laboratories Without Walls: Focused Precompetitive Research," *Trends in Biotechnology* 5:318-321, 1987.

DOE also engages, to a limited extent, in international research cooperation and collaboration through its national laboratories. It has been criticized, however, for earning "a poor reputation abroad for long-term commitment to international collaborations," which "will make it extremely difficult for DOE to attract foreign countries into significant new partnerships" (31). So far, however, DOE scientists working on genome projects have collaborated freely with researchers from other countries (82).

Existing research organizations can also become centers of collaboration. CEPH coordinates over 40 international investigators and research laboratories for mapping studies (see box 7-B). It sends genetic materials to major gene mapping laboratories around the world; in exchange, the laboratories share their results and data.

Washington University-RIKEN Collaboration

A recent agreement between researchers from Washington University in St. Louis, Missouri, and the Institute of Physical and Chemical Research (RIKEN) in Tsukuba, Japan, illustrates the potential of international collaboration at the level of individual institutions (38). The 3-year program, effective November 1, 1987, enables researchers from Washington University's new Center for Genetics in Medicine (founded by a donation from philanthropist James McDonnell) to work with researchers from the Tsukuba Life Sciences Center of RIKEN. The research will combine the expertise of the university's scientists in cloning yeast cells with the technological know-how of the RIKEN scientists, who have developed automated DNA analysis equipment. The initial focus of the research will be to sequence the entire yeast genome and to improve techniques for cloning human chromosomes into yeast cells.

This collaboration, the first bilateral agreement between American and Japanese scientists in the field of genetics, also provides for information and personnel exchanges with the Pasteur Institute in Paris and the Academia Sinica in Shanghai, China. Data and results from the collaboration will be disseminated freely to the international community.

International Human Gene Mapping Workshops

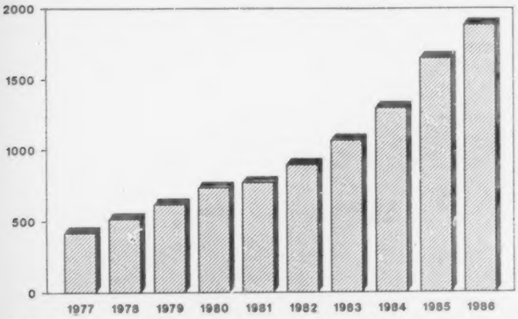
A series of biannual international gene mapping workshops—the ninth (HGM 9) was held in Paris in September 1987—has provided a mechanism for extensive international interaction. Prior to each workshop, committees are appointed for each of the human chromosomes. The committees are in charge of evaluating the research that has been done on the chromosome; they solicit papers from the international research community and select the ones to be presented. At the workshop, the committee for each particular chromosome works toward a consensus on which mapping data will be accepted as the standard. The committees also decide upon the official nomenclature for map sites and for probes, and their deliberations provide a measure of quality control for the research. Data accepted at the workshop are submitted to the Human Gene Mapping Library in New Haven, Connecticut, and subsequently entered into that database (see app. D). In 1987, a new database, Genatlas, was initiated specifically for the purpose of managing the mapping data from HGM 9. The conference proceedings are published in the *Journal of Cytogenetics and Cell Genetics*. Proceedings of some of the conferences have been independently published.

The growth in the size of the HGM workshops is one indication of the overall growth of the field of human genetics. Early conferences attracted an exclusive group of participants, but the ninth drew hundreds. Data are accumulating so rapidly that biannual conferences may not be sufficient; plans are already underway for an informal workshop, dubbed HGM 9.5, to be held in 1988 (9).

International Journals

The scientific publication process is the most important form of data sharing within and across national borders—an ongoing form of international cooperation. A bibliometric analysis of the international literature showed a rapid rise in the number of mapping and sequencing articles published in international journals between 1977 and 1986 (see figure 7-2). U.S. researchers have con-

Figure 7-2.—Human Gene Mapping and Sequencing Articles Published Annually



A bibliometric analysis conducted for the Office of Technology Assessment by Computer Horizons, Inc. [app. A] showed a steady increase in the total number of articles published annually in international journals on human gene mapping, gene markers, nucleotide sequences, and related topics from 1977 through 1986. [See app. E for details on the key words used in the literature search.]

SOURCE: Office of Technology Assessment, 1988.

sistently contributed the largest number—from 38 to 46 percent of all articles with genetic map or linkage results (see figure 7-1, table 7-1, and app. E) [Computer Horizons, Inc., see app. A]. The United Kingdom is the next largest contributor, publishing 8 to 11 percent of the articles annually, while France and West Germany are next with 5 to 8 percent and 2 to 5 percent, respectively. Japan's share of the basic research has increased fairly steadily, from 2 percent to 5 percent of the total. These data show the international nature of genome research and of the medical and scientific literature in general. There exists some segregation of Eastern European journals due to restrictions on export of information, and language may pose a barrier for non-English-speaking scientists (since many international journals are published in English), but for the most part scientific journals are thoroughly international. Scientists from one nation freely report data in journals from another.

Databases and Repositories

The operation of databases and repositories has been a standard mode of international cooperation in many scientific fields, and human genome projects are no exception; several databases and repositories relevant for human genome projects exist (see app. D). The cooperative arrangements that have evolved among the international databases for nucleotide sequences and for protein sequences are examples of effective international collaboration.

Databases for nucleotide sequences were started at Los Alamos National Laboratory (later funded by NIH and operated under the name GenBank®) and EMBL (officially dubbed the EMBL Data Library) during the late 1970s. By the fall of 1980, the database organizers recognized the need for collaboration between the two, and from 1980 through 1982 the databases exchanged sequence data on an informal basis until their first major releases. In August 1982, GenBank® and EMBL held their first joint meeting and agreed to use a similar system of accession numbers and to divide the journals each would scan for data. The compatibility of the databases was further enhanced by agreements, reached in 1985 and 1986, on common sets of data and annotation. The DNA Data Bank of Japan formally joined the collaboration in May 1987. The division of responsibilities for various aspects of the operation of the databases was formalized in meetings in the summer and fall of 1987.

An international workshop on database needs in molecular biology was convened in Heidelberg, West Germany, in 1987. The participants recommended that an international advisory committee composed of experts from the fields of molecular biology and information sciences be formed to provide advice and guidance for expanded cooperation among the databases (35). The funding agencies that support the databases followed the recommendation and appointed a committee, which consists of three members from the United States, three from Europe, and two from Japan. The committee will meet yearly to advise database staff on matters such as format and annotation. Its recommendations are not binding, however, since each database is responsive primarily to the agencies that support it. The first meeting was held in February 1988.

Formal collaboration on protein sequence databases is more recent. The U.S. database, the Protein Identification Resource (formerly known as the Dayhoff database, or NBRF), was started in the late 1960s. The European and Japanese counterparts—the Martinsreid Institute for Protein Sequence Data (MIPS) (27) and the Japan International Protein Information Database (JIPID)—began operations in 1987. The close collaboration among the three includes use of the same format, the same software, and a regional division of monitored journals (41).

The continued development and maintenance of databases and repositories are the most commonly endorsed mode of international cooperation on human genome projects (see box 7-D). The National Academy of Sciences supported the establishment of an international organization to gather and distribute data and materials in its 1988 report on human genome mapping (62).

CHAPTER 7 REFERENCES

1. *Abstracts in Biocommerce* 10:11, January 1988.
2. Altenmueller, G.H., " 'Genetic Engineering' Commission of Inquiry Presents Its Report: No Blinders, but Firmly in Hand: No Straitjacket for Genetic Engineering Research, Only Clearly Defined Guidelines," *VDI Nachricten* (Düsseldorf), Jan. 23, 1987, p. 1.
3. Ananyev, Y.V., Bochkanov, S.S., Ryzhik, M.V., et al., "Characterization of Cloned Repetitive DNA Sequences of Barley Genome," *Genetika* (Moscow) 22:920-928, 1986.
4. Anderson, A., "U.S. Pressures Japan Over Imbalance in Basic Research," *Nature* 329:662, 1987.
5. Anderson, M., Embassy of Finland, personal communication, December 1986.
6. Bolund, L., University of Aarhus, Denmark, personal communication, December 1987.
7. Borodovskiy, M.Y., Sprizhitskiy, Y.A., Golovanov, Y.I., et al., "Statistical Characteristics of Primary Functional Regions of *Escherichia Coli* Genome, Part 3, Computer Recognition of Coding Regions," *Molekulyarnaya Biologiya* (Moscow) 20:1390-1398, 1986.
8. Bowman, J., University of Chicago Medical School, personal communication, December 1987.
9. Cahill, G., Howard Hughes Medical Institute, personal communication, December 1987.
10. Cameron, C.M., Department of National Health and Population Development, South Africa, personal communication, October 1987.
11. Cantley, M., Commission of the European Economic Community, personal communications, August 1987; January 1988.
12. Centre d'Etude du Polymorphisme Humain, unpublished report, 1986.
13. Chamberlin, J.W., U.S. Embassy, Brazil, personal communication, August 1987.
14. CODATA, "First CODATA Workshop on Nucleic Acid and Protein Sequencing Data," program and abstracts from conference held at National Bureau of Standards, Gaithersburg, MD, May 3-6, 1987.
15. Collins, J., and Coulson, A.F.W., University of Edinburgh, personal communication, December 1987.
16. Commission of the European Communities, "Gene Technology: Opportunities and Risks," unpublished translation of the foreword and recommendations of a report by the Committee of Inquiry of the German Bundestag.
17. Commission of the European Communities, "Proposal for a Council Regulation on a Community Action in the Field of Information Technology and Telecommunications Applied to Health Care: AIM (Advanced Informatics in Medicine in Europe), Pilot Phase," COM (87) 352 final, Brussels, July 24, 1987.
18. Commission of the European Communities, "BICEPS Background Documents: Outline of a Community Action in the Field of Medical Informatics," European Research Organisation for Advanced Informatics in Medicine, XII/1036/86, August 1986.
19. Commission of the European Communities, "Single European Act," *Bulletin of the European Communities*, Supplement 2/86, 1986.
20. Commission of the European Communities, *Official Journal of the European Communities*, No. L 83/4, Mar. 25, 1985.
21. Conseil Européen des Fédérations de l'Industrie Chimique, "Bio-Informatics in Europe: An Industry Position Paper" (Brussels: CEFIC, 1987).
22. Consiglio Nazionale della Richerche, "Progetto Strategico C.N.R.: Mappaggio e Sequenzimento del Genoma Humano," unpublished manuscript, 1987.
23. Crawford, M., "Japan's U.S. R&D Role Widens, Begs Attention," *Science* 233:270-272, 1986.
24. Dahllof, S., "500 Million Kroner State Support to Danish Biotech," *NY Teknik* (Stockholm), Jan. 29, 1987, p. 7.
25. Dausset, J., "Human Polymorphism Study Center," paper presented at the International Conference on New Genetics and the Human Gene Map, Paris, Sept. 11-12, 1987.
26. Dausset, J., Cann, H.M., and Cohen, D., "Centre d'E-

tude du Polymorphisme Humain (CEPH): Collaborative Mapping of the Human Genome," unpublished manuscript, April 1987.

27. Dickman, S., "New Protein Database for Europe," *Nature* 327:265, 1987.
28. Dickson, D., "EMBL: 'Small Science' on a European Scale," *Science* 237:1108-1109, 1987.
29. Dickson, D., and Norman, C., "Science and Mutual Self-Interest," *Science* 237:1101-1102, 1987.
30. Dowdle, E.B., MEDGENE, South Africa, personal communication, October 1987.
31. Energy Research Advisory Board, *International Collaboration in the U.S. Department of Energy's Research and Development Programs*, DOE/S-0047 (Washington, DC: U.S. Government Printing Office, 1985), p. 2.
32. Epstein, J., OTA, personal communication, December 1987.
33. "ESF's Seibold on Forging Links for European Science," *The Scientist*, October 19, 1987, pp. 16-17.
34. "Eureka's Adolescent Growing Pains," *Le Figaro* (Paris), July 12, 1987, p. 21.
35. European Molecular Biology Laboratory and National Institutes of Health, "Future Databases for Molecular Biology," unpublished report from a workshop held Feb. 25-27, 1987, Heidelberg.
36. European Molecular Biology Laboratory, *The 1987-1990 Scientific Programme of the EMBL (SP87)*, EMBL/86/3, rev. 1 E, Sept. 1, 1986.
37. European Space Agency, *Memorandum of Understanding Between the European Space Agency and the United States National Aeronautics and Space Administration*, ESA/C(77)51, rev.1 Annexe, unpublished.
38. Fitzpatrick, T., "Washington University, Japanese Research Group Sign Genetic Research Agreement," news release, Washington University, St. Louis, MO, Dec. 7, 1987.
39. Fujimura, R.K., *Biotechnology in Japan* (Washington, DC: Department of Commerce, in press).
40. Fujimura, R.K., Oak Ridge National Laboratory, personal communications, August 1987; December 1987.
41. George, D., Protein Identification Resource, Washington, DC, personal communication, December 1987.
42. "Genome Analysis Can Increase Therapy Chances for Genetically Determined Diseases," *Handelsblatt* (Düsseldorf), May 14, 1987, p. 6.
43. Gibson, F., Australian National University, personal communication, January 1987.
44. Gilbert, W., "Genome Sequencing: Creating a New Biology for the Twenty-First Century," *Issues in Science and Technology* (spring): pp. 26-35, 1987.
45. Grienitz, H., "Human Genetics—Humane Genetics," *Spectrum* (East Berlin) 3:10-13, 1987.
46. "Human Frontier Wins Muted Support," *New Scientist*, June 18, 1987, p. 32.
47. Hunkapillar, M., Applied Biosystems, Inc., personal communication, October 1987.
48. Ingman, B.J., "Finland: 184 Million Markka Biotechnology Program 1988-1992," *Hufvudstadsbladet* (Helsinki), January 31, 1987, p. 15.
49. Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
50. Jenkins, T., University of Witwatersrand, Johannesburg, personal communication, November 1986.
51. Kaminsky, H., U.S. Embassy, Italy, personal communication, July 1987.
52. Klemenz, H., University of Heidelberg, personal communication, September 1987.
53. Kohara, Y., Akiyama, K., and Isono, K., "The Physical Map of the Whole *E. Coli* Chromosome: Application of a New Strategy for Rapid Analysis and Sorting of a Large Genomic Library," *Cell* 50:495-508, 1987.
54. Logsdon, J., George Washington University, personal communication, August 1987.
55. Logsdon, J., "U.S.-European Cooperation in Space Science: A 25-Year Perspective," *Science* 223:11-16, 1984.
56. Maddox, J., "Brenner Homes in on the Human Genome," *Nature* 326:119, 1987.
57. Mannarino, E., Embassy of Italy, personal communication, July 1987.
58. Matsubara, K., Institute for Molecular and Cellular Biology, Osaka University, Japan, personal communication, February 1987.
59. Mohr, J., University of Copenhagen, personal communication, August 1987.
60. Morris, R.G., U.S. Embassy, Argentina, personal communication, September 1987.
61. National Aeronautics and Space Administration, *Life Sciences Report, 1987* (Washington, DC: NASA, 1987), p. 65.
62. National Research Council, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).
63. Novello, A.C., "National Institutes of Health Awards to Institutions in Foreign Countries, 1976-85," *The Lancet*, Sept. 6, 1986, pp. 561-563,
64. Oddo, G., "More Business Involvement in CNR's Finalized Programs," *Il Sole 24-Ore* (Milan), July 14, 1987, p. 6.
65. Organización de los Estados Americanos, *Situacion de la Genetica en la Region: Informe Preliminar* (Washington, DC: OAS, 1986).

66. Probert, M., Medical Research Council, United Kingdom, personal communications, July 1987; December 1987.
67. Roe, L., Ministry of State, Canada, personal communication, December 1987.
68. Schmitz, D., "Biotechnology in the Federal Republic of Germany (FRG)," report prepared for the U.S. Department of State, December 1986.
69. Simpkins, L.C., U.S. Embassy, Mexico, personal communication, September 1987.
70. Skou, B., Royal Danish Embassy, personal communications, November 1986; February 1987; April 1987.
71. Smith, R., The Johns Hopkins University, personal communication, July 1987.
72. Soll, D., Yale University, personal communications, August 1987; September 1987.
73. Sommariva, A., "And Italy Will Study Chromosome 22," *Italia Oggi* (Milan), May 22, 1987, p. 36.
74. Sukholdolets, V.V., "Bacterial Genome Construction: New Advances in Genetic Engineering," *Genetika* (Moscow) 22:901-903, 1986.
75. Swinbanks, D., "Human Frontiers Program Seeks International Help," *Nature* 330:683, 1987.
76. Swinbanks, D., "Japan's Human Frontiers Project Stays in the Doldrums," *Nature* 328:100, 1987.
77. Swinbanks, D., "What Future Now for Japanese Biotechnology Research?" *Nature* 329:661, 1987.
78. Technologie Elettriche (Milan), January 1987, pp. 78-87.
79. U.S. Congress, Office of Technology Assessment, *Commercial Biotechnology: An International Analysis* (Springfield, VA: National Technical Information Services, 1984).
80. U.S. Congress, Office of Technology Assessment, *Star Power: The U.S. and the International Quest for Fusion Energy*, OTA-E-338 (Washington, DC: U.S. Government Printing Office, October 1987).
81. U.S. Congress, Office of Technology Assessment, *New Developments in Biotechnology, 4: U.S. Investment in Biotechnology* (Washington, DC: U.S. Government Printing Office, in press).
82. Van Belkom, P., National Health and Medical Research Council, Australia, personal communications, December 1986; November 1987.
83. van der Meer, R.R., "EC-biotechnology: European Challenge," *Trends in Biotechnology* 4:277-279, 1986.
84. Wada, A., "Japanese Super DNA Sequencer Project," *Science and Technology in Japan* 6 (22):20-21, 1987.
85. Wada, A., "Automated High-Speed DNA Sequencing," *Nature* 325:771-772, 1987.
86. Wada, A., University of Tokyo, personal communication, September 1987.
87. Wada, A., "Strategy for Building an Automatic and High Speed DNA-Sequencing System," in *Proceedings of the 4th Congress of the Federation of Asian and Oceanian Biochemists* (London: Cambridge University Press, in press).
88. Yarrow, D.J., British Embassy, personal communications, January 1987; December 1987.
89. Yoder, S., "Japanese Tap Advanced Technologies To Accelerate Deciphering of DNA," *Asian Wall Street Weekly*, Dec. 14, 1987, pp. 1, 24.
90. Yoshikawa, A., University of California, Berkeley, personal communication, December 1987.
91. Yuan, R., *Biotechnology in Western Europe* (Washington, DC: Department of Commerce, 1987).

Chapter 8

Technology Transfer

CONTENTS

Page

Patent and Copyright Policies 166
  Patents 166
  Copyrights 170
  Trade Secrets 170
International Technology Transfer 172
  Economic Benefits 172
  Humanitarian and Scientific Benefits 173
  National Prestige 174
  Military Applications 174
Chapter 8 References 174

Chapter 8

Technology Transfer

> "The politics of knowledge—the question of who owns and controls the distribution and use of scientific information—is by no means a new issue. The pure scientist working in an ivory tower has long been extinct."
>
> Dorothy Nelkin, *Science as Intellectual Property: Who Controls Research?* (Washington, DC: American Association for the Advancement of Science, 1984), p. 92.

The economic impact of genome projects will depend on how many new products and services are created by them. Some large scientific projects such as space programs and electronics research facilities have been justified by their potential for spinning off technologies. The magnitude of such spinoffs is unpredictable, however. Often, there emerge useful products that could not have been foreseen [Heilbron and Kevles, see app. A]. Given the many surprises in molecular biology over the past decade, it is impossible to predict exactly how genome projects will result in products, but they undoubtedly will yield many new applications in pharmaceuticals, agriculture, and other industrial sectors. Uncertainty about the magnitude of economic impact means that genome projects cannot be justified purely as an economic investment. As the projects go forward for scientific and medical reasons, however, it makes sense to ensure that their results are fully used. The process of converting scientific knowledge into useful products is *technology transfer*.

The Federal Government influences the efficiency of technology transfer through its research and development policies. Government has traditionally supported research that will have large but unmeasurable noneconomic benefits (e.g., research aimed at improving health as a value in itself rather than simply disease impact measured in dollars) or that is too risky for individual firms to support (e.g., projects that are expensive, highly uncertain in outcome, or long-term). Arguments for increased Federal support of biomedical research since World War II have generally emphasized improvements in health. Economic arguments for increased biomedical research funding have typically been analyses of economic drag—how much the Nation could save by avoiding disability or disease (18). This argument is changing to concern for efficient translation of science into products. Policymakers are shifting their attention to technology transfer as products derived from molecular genetics find their way to the marketplace, international trade imbalances worsen, and rising deficits intensify scrutiny of Federal budgets.

A major effort is underway in many developed and some developing nations to target biotechnology for investment because it is considered particularly likely to produce economic benefits (3,16,19,23). Most foreign governments' efforts to promote biotechnology include strategic planning of national research programs and encouragement of research and development in private firms (e.g., tax incentives, subsidies for industrial research centers, business grants, or government risk capital). The United States has no deliberate Federal policy to encourage biotechnology per se (16,19,23), although legislation introduced late in the first session of the 100th Congress would create a national biotechnology policy board.

Most genome projects could produce both direct and indirect economic benefits. Some projects are expected to yield directly marketable products (e.g., DNA sequenators, analytical instruments, DNA probes for diagnostic tests). Others would accelerate development of products (e.g., maps, repositories, and databases).

Different groups have divergent concerns about technology transfer. Scientists fear that corporate participation will inhibit the free flow of information and impede scientific progress. Policymakers want to ensure that a large Federal investment in genome projects is translated efficiently into new products and services, ultimately creating new jobs and other economic benefits. They are wary of projects in which U.S. taxpayers will fund research that is commercialized and used by foreign interests. In this view, foreign governments should support an equitable fraction of basic research, and American investments should not allow jobs and profits to migrate abroad. Industrial representatives want a say in planning research programs and access to scientific results as they are produced. Individual companies wish to ensure that any funds they invest will earn sufficient returns.

Congress could encourage technology transfer by funding personnel exchange among government, academic, and industrial sectors, with minimal bureaucratic strictures, and by supporting symposia, journals, and other modes of information exchange. When advisory committees are formed to guide Federal genome projects, industrial representatives could ensure that projects are planned with an eye to economic exploitation. These options are covered in chapter 6. The remaining options relate to protection of inventions resulting from federally funded research, discussed below.

PATENT AND COPYRIGHT POLICIES

Ideas and know-how—intellectual property—are granted many of the same legal protections as tangible private property. Intellectual property law traces its roots directly to the U.S. Constitution, which authorizes the Federal Government "to promote the Progress of Science and the useful Arts, by securing for limited times to their Authors and Inventors the exclusive Right to their respective Writings and Discoveries." The purpose of intellectual property protection is to encourage inventors and discoverers to share their knowledge, while ensuring that they benefit from the fruits of their labors. Legal protections balance the social good stemming from wide disclosure of new knowledge against individuals' or companies' rights to gain from what would not have existed without their efforts.

Three types of intellectual property protection are relevant to discussion of the technologies likely to emerge from genome projects: patents, copyrights, and trade secrets.

Patents

Patents grant inventors the right to exclude others from producing, using, or marketing their inventions (as defined in the patent claims) for a specified period. The purpose of patent law is to give inventors an incentive to risk their time and money in research and development, while requiring public disclosure. Patent laws in different countries vary in degree of protection, enforcement, penalties for violation, and criteria for approval. In the United States, the period of protection is 17 years, with extensions for pharmaceuticals to cover some of the delay imposed by regulation. Patents apply to inventions, but not to ideas, mathematical formulas, or discoveries of preexisting things. A patentable invention must be new, useful, and not obvious. A patent holder can permit others to use or make the invention by licensing it.

Profit is only one of many motivations for patenting an invention. Another is to maintain control over it. Leo Szilard filed a patent on the process of nuclear fission, for example, hoping to bring it to the attention of military authorities in the United States and Great Britain (12). Cyclotrons used in nuclear physics were patented to ensure their proper medical applications, yet this did not inhibit research (in fact, most physicists were not even aware of the patents) [Heilbron and Kevles, see app. A]. The Rockefeller Institute patented the sphygmomanometer (blood pressure cuff) to ensure that clinicians would have ready access to it and that later discoverers could not limit its use (11). Nonprofit organizations supporting genome projects are likely to encourage patents when they would ensure broad public use (9).

U.S. patents are obtained from the Patent and Trademark Office in the Department of Commerce (other nations have analogous institutions). The patentability of inventions is initially determined by this office. The scope of protection and more refined factors for granting patents are defined by case law, when patents are challenged in court. The principles for determining patentability do not depend on any particular type of technology, but interpretation of them does. Uncertainty about the patentability of inventions is greater in biotechnology than in most other areas because the techniques are new and complex. Interpretation of criteria for granting patents, defining the scope of patent claims, and determining what constitutes infringement have not been clarified by case law, because the case law does not yet exist.

Patent Policies for Federally Funded Research

Uncertainty about patentability need not paralyze research efforts, because interpretation of patent law does not interfere with most federally supported research (as explained below). Patent policies of Federal agencies will nonetheless influence how genome research is commercialized, and these patent policies have changed dramatically over the past decade. The changes are intended to promote commercial application of federally funded research by permitting private ownership and control of its results. The reasoning is that research will be more broadly disseminated and effectively used if those who conduct it are granted title to the patents on resulting inventions, thus providing an incentive to commercialize the inventions [Rosenfeld, see app. A].

Changes in patent policy resulted from studies showing that, while the Federal Government held title to roughly 28,000 inventions in 1975, fewer than 5 percent had been licensed to businesses (15). The Patent and Trademarks Amendments of 1980 (Public Law 96-517) were passed to grant title to small businesses and nonprofit organizations funded to do research by the Federal Government. These were further amended in the Trademark Clarification Act of 1984 (Public Law 98-620), most significantly by removing restrictions on licensing. Regulations implementing these laws were made final by the Department of Commerce in March 1987.

The policies applying to small businesses and nonprofit organizations were extended to large businesses, with some exceptions, by a memorandum from President Ronald Reagan dated February 18, 1983. The Technology Transfer Act of 1986 (Public Law 99-502) permitted new licensing and joint venture arrangements, and granted agencies authority to form consortia with private concerns. Executive Order 12591, issued by President Reagan in April 1987, encouraged technology transfer of federally funded research. The order was based on existing statutes and promoted consortium formation, exchange of research personnel between government laboratories and industrial firms, special technology transfer programs at federally owned laboratories, and transfer of patent rights to government grantees and contractors.

The policies embodied in these statutes, regulations, and executive orders constrain the authority of Federal agencies to force sharing of data if sharing would conflict with recipients' taking title to inventions [Rosenfeld, see app. A]. The government can override recipients and take title to patents only in special situations. One of these is when an agency determines that retaining title "will better promote the policy and objectives" of the patent statutes. This clause has been narrowly interpreted and has rarely been used by the research agencies involved in genome projects [Rosenfeld, see app. A]. The Federal Government can also impose licensing requirements to "alleviate health and safety needs," "meet requirements for public use specified by Federal regulations," or meet "certain statutory provisions requiring products to be manufactured in the United States." These provisions have also been narrowly interpreted and impose such a daunting burden of proof on agencies that they are unlikely to be used. They could conceivably be invoked if patent rights interfered with the pooling of data that must be collective to be useful or if clinical benefits were delayed (e.g., slow commercialization of genetic tests or therapies), but only if problems were severe and obvious.

Federal research agencies' patent policies need not unduly slow exchange of information. The degree to which information flow is impeded will depend on when grant recipients and contractors file patent applications. Many genome projects will result in patentable inventions, particularly those focused on technology development. Recipients of Federal funds may follow one of three courses of action: file applications early and subsequently release data; file early and do not take extra actions to release data (relying on the patent process to do so); or decide not to patent.

Filing patent applications early and publishing data soon thereafter are optimal for encouraging rapid dissemination of knowledge, protecting inventors' rights, and preserving economic benefits in the United States. Early patenting and subsequent disclosure would release data for public use but would help inventors maintain control of their inventions and assure them and their sponsoring institutions of any financial rewards. Early patent application would also protect the Nation because statutes give a preference to U.S. manufacture of any resulting products or services. Early patent application not followed by special efforts to disseminate data would ensure benefits for the grant recipient or contractor but would needlessly delay exchange of useful information—patents are often not published for several years, and it has taken over 7 years for some biotechnology patents to be awarded.

Investigators may decide not to apply for a patent because they wish to avoid substantial legal costs and bureaucratic entanglements or because they believe that science should not become commercially oriented. This can make new methods freely available to all, but it can also inhibit full exploitation of an invention. It is also against the intent of Federal statutes, which require recipients of Federal funds to report patentable inventions. An inventor can lose control of an invention if he or she does not file a patent and another inventor does so. A product or process that is not patented is unlikely to be used commercially, because any firm investing in manufacture will want a guarantee that its investment will be protected. Failure to patent also invites foreign exploitation of research funded at U.S. taxpayers' expense: Patent rights could be claimed by a foreign company, research institution, or individual; U.S. firms would not be given manufacturing preference; and the U.S. inventor could be prevented from use of the invention. Export of economic benefits has occurred frequently in biological sciences when initial discoveries have not been patented. Penicillin was discovered in England, for example, but the patent was obtained by U.S. corporations. The cell fusion process for making monoclonal antibodies was developed in London, but many of its applications were exploited first in the United States. In both cases, the United Kingdom claimed the Nobel Prize, but the United States reaped most of the economic benefits.

Federal agencies and Congress may wish to oversee patent practices of grantees and contractors closely to ensure that patents are filed early and data exchanged soon thereafter. Disclosure of *data* should not be long delayed by policies designed to encourage patenting of inventions, because data per se are not inventions eligible for patent protection. There is a gray area, however, between invention of new methods and the data that result from using them.

Scientists may be reticent to disclose details of methods used to generate data if doing so endangers patentability. An invention must be novel to be patented: that is, it must not be widely used by parties other than the inventor for more than one year, and publication of the method cannot precede filing the patent by more than one year. (Some foreign countries do not permit even the one-year grace period.) If investigators are uncertain whether disclosing details of method would threaten a patent, they may choose not to publish those details. Uncertainty over patentability can indeed inhibit the free exchange of information. It has led one commentator to list three possible ways of altering patent laws: 1) making the definition of novelty more flexible; 2) establishing an intellectual property protection that is analogous to but more limited than patents and that requires less rigorous proof of novelty and nonobviousness; or 3) legislating special intellectual property protections for biotechnology (8). Further study is needed "to determine whether and how biotechnology demands special treatment as intellectual property before legislative reform will be in order" (8). This suggests that patent policies might be high on the agenda for congressional oversight but low on the legislative calendar.

Filing patents early and then disclosing the results could worsen an already considerable backlog of pending patents. Approximately 7,000 biotechnology patents have been filed at the Patent and Trademark Office and await final action (20). If the benefits of patent protection are judged important by Congress, then one option would be to increase the resources in the biotechnology sections of the Patent and Trademark Office. This could include higher salaries, more opportunities for training to keep abreast of technological developments, easier access to technical databases, and more examiners. Increased resources could not only reduce uncertainty by diminishing the backlog of pending patents, but also increase the attention devoted to each application and reduce subsequent litigation.

Patent Policies at Research Agencies

The Department of Commerce recently promulgated final regulations for Federal agencies to use when funding research at small businesses, universities, and nonprofit organizations [37 CFR §§401]. While these regulations, issued in March 1987, have had little time to take effect, the National Science Foundation (NSF) and the National Institutes of Health (NIH) have followed similar policies since the late 1970s.

The General Accounting Office found that university administrators, industry representatives, and small businesses all reported a "significant positive impact on research and innovation" from taking title to inventions that resulted from federally funded research. University and industry officials also reported benefits from the 1984 law that removed licensing restrictions (15). Agencies likewise reported a generally positive assessment, with greater potential for licensing patents than when title was retained by the Federal Government.

The situation at the Department of Energy (DOE) is more complex. A substantial fraction of DOE research funding goes to national laboratories, which are owned by the Federal Government and operated by private contractors. At most of the laboratories, the contractor can elect to take title to inventions. Title rights are restricted, however, at facilities that conduct research on weapons and naval propulsion systems. This could prove relevant to genome projects because several of the groups that have been directly engaged in DOE's Human Genome Initiative are located at laboratories with restricted title policies—namely, Lawrence Livermore National Laboratory and Los Alamos National Laboratory, both operated by the University of California. Regulations state that limitations on the contractor's right to take title should be restricted to "inventions occurring under" naval nuclear propulsion or weapons-related programs [37 CFR Part 401.3(a)(4)]. This should permit the contractor to take title to inventions from human genome projects because the projects would not be conducted under restricted programs, even at the affected laboratories. Negotiations between DOE and contractors are more complicated, however, when restrictions differ among programs at the same facility. Legislation has been proposed to mandate patent policies for genome projects at the national laboratories; the policies would be modeled on those of other research agencies.

The regulations and executive orders implementing patent policies at research agencies are quite recent. It would be premature to alter those policies fundamentally until the results of current law can be assessed (with the possible exception of DOE policies regarding national laboratories, noted above).

There are additional roles for Congress. First, Congress could monitor the practices of Federal agencies and funding recipients to ensure that the intent of existing statutes is carried out. Second, Congress could increase resources to the Patent and Trademark Office to enable more efficient processing of patents. Third, Congress could increase resources for universities and other recipients in order to manage patent filing in the United States and abroad. Finally, Congress could ask agencies engaged in genome projects to specify their patent policies more clearly. At present, written material on patent policies at NIH, DOE, and NSF is difficult to obtain, and there is no single source for information on patent policies at all agencies involved in genome projects [Rosenfeld, see app. A]. The interagency nature of genome projects means that recipient institutions will often be funded by more than one agency. A clear presentation of patent guidelines at various agencies, with explanations of the advantages of early patent filing and the implications of doing so (and not doing so), might diminish confusion and promote commercial application.

Copyrights

Copyright law is intended to protect works of authorship. It has traditionally been applied to works of art, books, and articles but has had to adapt to technological change. Copyrights now extend to computer software and electronic entertainment media, for example (6,17). Copyright is intended to protect the *expression* of ideas, not the ideas themselves—a difficult but crucial distinction.

The Copyright Act of 1976 is the most recent statute relevant to genome projects, extending protections to nontraditional media such as computer software. The extensions may also prove relevant for research in molecular biology (6). Case law has evolved doctrines to test the distinction between idea and expression and to define the scope of protection. An author can prohibit others from copying his or her book, for example, but the concepts and methods described in the book are not protected. Arguments have been made that copyright could apply to DNA (6), but this line of argument is not widely accepted and the scope of protection (if it exists) is quite narrow (5). The ability to copyright a native DNA sequence derived from a human chromosome or other natural source is particularly uncertain (5). A preliminary communication from the Copyright Registration Office indicates that such sequences would not be accepted, although the book or printed map containing them—the particular expression of map or sequence data—would (10).

Even if DNA maps can be copyrighted, such copyrights are unlikely to inhibit research substantially. In normal circumstances, obtaining a copyright does not require extra time and is thus not a justification for delaying disclosure of results. A company could charge for access to map or sequence information in much the same way that commercial databases charge for information sharing. Access and service charges are not new—molecular biologists routinely pay for services that are less expensively or more rapidly performed by others. They buy copyrighted books and read copyrighted journals. Many materials used in biological research (clones, enzymes, chemicals) can be made by individual investigators, but it is easier to purchase such materials from a company set up to make them.

The type of research conducted by a private company engaged in mapping and sequencing DNA would be feasible in a large number of laboratories. Copyrights would not prevent investigators from using *information* published or otherwise provided by a company or from duplicating the work. A company that has developed extensive map and sequence information would either charge so little that it is cheaper for a researcher to obtain it from the company than to do the work, or the researcher would in fact repeat the work. In either case, the community of researchers is no worse off than if the company had not mapped or sequenced.

If copyright practices prove to impede research, then agencies can take steps to correct the deficiencies. Agencies have much broader discretion for copyright policies than for patents [Rosenfeld, see app. A].

Trade Secrets

Information held by one company that is useful in its business and unavailable to competitors is called a trade secret. Trade secrets can be protected from misappropriation—that is, improper disclosure—through the courts, which award monetary damages for unauthorized use. A trade secret must be in continual use, be well established in practice, and have actual or potential commercial value (19). The holder must take steps to guard it. Trade secrets do not involve slow and costly legal steps for registration, their duration is not limited by law, and they need not meet patent or copyright criteria. Uncertainties about patents and copyrights are not relevant (although legal criteria for protection under trade secret laws must be met). Trade secret protections are principally secured under State rather than Federal laws, and there is some variation among the States. Trade secrets have limited scope: In a rapidly moving field they may not last long. Trade secret laws cannot ensure returns on a research investment if another inventor discovers the secret method or finds a new way to do the same thing. Protection does not apply, even if competitors figure out the secret by examining a product (reverse engineering). Most important, trade secrets must be kept secret. This would be quite difficult to justify for federally funded research.

The scientific equivalent of a trade secret is nondisclosure. This is referred to pejoratively as sitting on data and is widely viewed as improper beyond the period needed to confirm accuracy of results and take advantage of a lead for further research. The period of nondisclosure varies widely among researchers, even those in the same field. Researchers who share data and materials early and freely are widely praised, such as the many collaborators who worked to find the muscular dystrophy gene (see ch. 3). But nondisclosure—for a few months to a year—is not uncommon in order to maintain a research advantage or to establish first discovery, even in research leading to Nobel Prizes (or perhaps especially in such research) (21,22). Permanent nondisclosure of an important result is, however, inimical to the purpose of scientific inquiry—the discovery and dissemination of new knowledge.

Nondisclosure is of particular concern when the results must be pooled in order to be useful (e.g., maps derived from data contributed by various groups). The need for pooled data can create a situation known as the prisoner's dilemma: when cooperation of all parties yields the maximum benefits, but one party can benefit if he does not cooperate and the others do. (So called because prisoners planning a jail break all benefit from cooperation, but one stooge can benefit individually by telling the guard of the plans.) An investigator searching for the location of an unknown gene stands to gain if other groups with markers make them freely available but he does not. He can then use both his and others' work to speed the search, while denying others access to his markers. Similar situations will arise in connection with submitting information to databases, sending materials to other researchers or central repositories, and other cases directly related to genome projects. Agencies will need to monitor the free exchange of data and materials, particularly when the efforts must be collective, and take steps to correct inequities. The need for joint efforts highlights the importance and fragility of collaborative institutions such as the Center for the Study of Human Polymorphism (CEPH) (see ch. 7).

Many journals have either explicit or unwritten policies that research data and materials described in an article must be made available to other researchers at the time of publication. Researchers preserve their option for exclusive use from the time of discovery until publication. Many scientists make materials available even before publication, which can require many months. Linking availability of materials to publication is a powerful mechanism, because one measure of scientific prestige is priority—who discovered something first. Priority is generally determined by date of publication. In large collaborative scientific projects, mechanisms have evolved to permit scientists time to pursue hot research leads while ensuring that others gain fair access. (CEPH's policy of sharing one set of data only among collaborators and making another set publicly available is an example.)

An informal policy of disclosure operates in Federal agencies through the process of peer review. If a researcher is known to hoard data—and such information spreads rapidly through scientific communities—then proposals submitted by that individual are unlikely to be given high priority by study sections (7). Review groups withhold support from research whose results they cannot see. This mechanism is slow—it can only be used when a grant is up for renewal, every 3 years or more—but it can be quite effective. If further measures are needed, Federal agencies could require submission of materials and data—map positions or DNA sequence data, for example—to the appropriate database. Such a policy would not be easily enforceable, however, and would be constrained by investigators' patent rights. Some journals now require submission of DNA sequences in proper form to GenBank® or its sister database in Europe at the time a paper is accepted. Agencies could devise incentives to make contribution of data and materials attractive, an alternative that is more easily implemented and less politically troublesome than negative sanctions. Those submitting data to CEPH, for example, benefit from knowing the position of their markers relative to markers found by others. Persons managing the DNA sequence databases have contemplated giving researchers a similar incentive.

Federal agencies have substantial power to require disclosure when it does not impede grantees' and contractors' intellectual property rights. Grant recipients and contractors need ample time to file patent applications, but legal protections of intellectual property are unlikely to inhibit agency policies promoting disclosure, particularly when broad access to data is necessary to fulfill the agency's mission.

INTERNATIONAL TECHNOLOGY TRANSFER

Human gene mapping is inherently international in scope. Recent breakthroughs in assembling rough genetic maps, for example, have depended on an international collaboration of investigators from Europe, North America, and Africa using family data from four continents. Several current technologies for sequencing and physical mapping were developed in the United Kingdom and other European nations, not the United States; however, recent years have seen increased emphasis on retaining the economic benefits of federally funded research for the United States.

International technology transfer is the movement of inventions and know-how across national borders. Concerns about international technology transfer fall into four areas: economic benefits, humanitarian and scientific benefits, national prestige, and military applications.

Economic Benefits

Concerns about economic implications of international technology transfer focus primarily on the export of jobs and services generated by research funded at public expense. Policies to combat this fall into three main areas: patent policies, restrictions on flow of information and materials, and promotion of domestic technology transfer so that benefits remain within national borders.

The patent policies described above have several provisions on international technology transfer that are relevant to genome projects. For foreign recipients of Federal funds or those subject to a foreign government, agencies must consider whether the recipient's government or company enters into international cooperative funding agreements on a "comparable basis" and whether the recipient's government protects U.S. intellectual property rights [Executive Order 12591, Apr. 10, 1987]. Recipients of Federal R&D funds must ensure that the products of the invention will be "manufactured substantially in the United States" [35 U.S.C. §§204]. Since jobs and economic wealth are linked more tightly to manufacturing than to initial research and development, even foreign-held U.S. patents resulting from Federal funding would have economic benefits in the United States. Moreover, Federal agencies are not required to grant patent rights to foreign recipients or those subject to control of a foreign government, even if they are universities or nonprofit organizations [37 CFR 401.14(a)(1)]. Foreign recipients are thus managed differently than their U.S. counterparts. Agencies could conceivably require foreign recipients to assign title to the U.S. Government or require that U.S. research partners take title.

Exploiting federally funded research inventions abroad will usually entail seeking foreign patents. Several international conventions govern patents, but conditions for granting patents differ among nations. The United States permits a grace period of one year from the date of publication to file a patent application, for example, but many other governments do not. If investigators wish to ensure worldwide patentability, therefore, they must file foreign patents before publication. The period of patent protection also differs. Researcher institutions accepting Federal funds must know about these and other differences when making decisions about foreign patents. Disseminating knowledge about such differences could be encouraged by research agencies in concert with the Department of Commerce. Agencies could also encourage institutions receiving Federal funds to pursue foreign patents.

The current necessity for filing patents individually in many countries is expensive and wasteful for all nations. International patent policies have been discussed several times at meetings of the Organization for Economic Cooperation and Development. Attempts are being made to harmonize international practices (14).

Humanitarian and Scientific Benefits

The humanitarian and scientific benefits of genome projects will be great. The United States has consistently performed a significantly higher fraction of the total mapping and sequencing effort than any other nation (see ch. 7). The knowledge resulting from these efforts has been freely shared with the rest of the world, to the benefit of citizens of all nations. The scientific knowledge generated at Federal expense since World War II may well prove to be one of the most significant international contributions of modern American culture.

Imposing restrictions on the flow of information and scientific materials from U.S. researchers to researchers abroad would be politically troublesome and technically difficult. Details of what to share and what to restrict would be difficult to describe in advance, and policies restricting the flow of data are against scientific traditions, which transcend national borders. Withholding map locations and DNA sequence information would be a violation of scientific ideals, particularly when such information could be clinically useful. Unilateral restrictions imposed by the United States would invite reciprocation, to the detriment of worldwide scientific progress.

The same tradition of free international exchange does not necessarily apply to the exchange of services and products—for example, mapping services, instruments, automation equipment, and reagents—which is governed more by international trade agreements than by scientific practices. Many national governments wish to assist their companies in developing goods and services for export. Genome projects focused on technology development are likely to be seen in this light. Nationalistic economic policies make projects to develop instruments or other salable goods poor candidates for international cooperation. European nations may be exceptions, because they have a basis for cooperation through several biotechnology programs of the European Economic Community.

Restrictions on international exchange of scientific personnel would disrupt many molecular biology laboratories in the United States and abroad. The United States has often reaped the benefits of international scientific exchange. Senior scientists, postdoctoral fellows, and graduate students from other nations work in U.S. laboratories and attend conferences. In exchange, U.S. scientists visit and are occasionally educated at universities and research centers abroad (12,22). The team of scientists that developed the atomic bomb for the U.S. Army, for example, was heavily dependent on scientists trained in Europe (12). Molecular biologists from abroad have often settled in the United States because it is so conducive to scientific research; several Nobel laureates at American universities immigrated during their scientific careers. Many projects in molecular biology have depended heavily on foreign scientists working in the United States, and many of the best stay or eventually return (4). The United States may in fact benefit from international personnel exchanges more than it is hurt by them. The Federal Government could nonetheless limit funding of foreign researchers at U.S. institutions, although this would probably generate ill will and provoke reciprocal actions by other governments.

One of the problems in assessing the potential impact of policies to reduce funding of foreign researchers in American laboratories is the absence of information about their research careers. If most foreign researchers remain in the United States or are particularly productive investigators while receiving Federal funds, then policies to restrict their ingress would be counterproductive.

Extending current restrictions or use of Federal funds for American researchers to travel abroad would be even less politically acceptable and more difficult to implement. It would result in direct loss of information to the United States, because persons traveling abroad are as likely to import information from their foreign collaborators as to export it. Policies designed to inhibit the exchange of personnel, materials, and information across national borders threaten benefits but gain little for the United States.

Promoting domestic exploitation and foreign patenting of new technologies is a more positive and less politically troublesome means to the same end of improving U.S. economic competitiveness. Such policies can preserve the U.S. lead in research without provoking retaliation or tarnishing the country's prestige.

National Prestige

One argument for Federal sponsorship for genome projects is that they are highly conspicuous and beneficial: Other nations will do the work if the United States does not, to the detriment of U.S. prestige. Similar arguments have been proffered for the supersonic transport, space programs, and other technical projects. These arguments tie the stature of U.S. science and technology to leadership of genome projects. The international prestige attached to genome projects is a purely political judgment; it cannot be assessed technically.

What would be the consequences if Japan or a European nation were to have the first complete set of ordered DNA clones representing all human chromosomes, or the first reference sequence of the human genome? Such questions are best answered primarily by the scientific and technical merits of the projects, not by an appeal to a vague notion of national prestige. If projects are technically unsound or uneconomical, then the United States would not benefit from a commitment to them. Other countries could do so, but they would only be hurting themselves. If the projects are technically sound, then the United States would do well to lead or at least participate in them, but national prestige would not be the principal justification for involvement. National prestige is not a useful basis for judging major scientific or technical projects.

Military Applications

Military applications of results of genome projects should not prove to be a major consideration in technology transfer. U.S. policies ban the export of goods and technologies that could be used for military purposes by specified hostile countries. Such policies are administered by the Department of Commerce in consultation with the Department of Defense and the Department of State. The export of some goods produced using biological technologies could be affected (1). At present, however, DNA mapping, sequencing, and other means of analysis relevant to genome projects are not on the list of controlled technologies, and this should remain true for the foreseeable future (2). The main reason is that the technologies and data resulting from genome projects would not have immediate military applications. Like other technologies and data, some could conceivably be used for a military purpose, such as devising vaccines against biological warfare agents, but genome projects would not in themselves promote biological warfare.

CHAPTER 8 REFERENCES

1. Dashiell, T.R., "The Department of Defense and Biotechnology," *Technology in Society* 8:223-228, 1986.
2. Dashiell, T.R., comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
3. Fujimara, R.K., International Trade Administration, U.S. Department of Commerce, *Biotechnology in Japan* (Washington, DC: U.S. Government Printing Office, in press).
4. Hall, S.S., *Invisible Frontiers: The Race To Synthesize a Human Gene* (New York, NY: Atlantic Monthly Press, 1987).
5. Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
6. Kayton, I., "Copyright in Living Genetically Engineered Works," *George Washington Law Review* 50:191-218, 1982.
7. Kirschstein, R, comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
8. Korn, D., "Patent and Trade Secret Protection in University-Industry Research Relationships in Biotechnology," *Harvard Journal on Legislation* 24(winter):191-238, 1987.
9. Kumin, L., Howard Hughes Medical Institute, Bethesda, MD, personal communication, December 1987.
10. Library of Congress, Copyright Registration Office, Washington, DC, personal communication, May 1987.

11. Lowrance, W., comments at Issues of Collaboration for Human Genome Projects, OTA workshop, June 26, 1987.
12. Rhodes, R., *The Making of the Atomic Bomb* (New York, NY: Basic Books, 1986).
13. Rycroft, R.W., "International Cooperation in Science Policy: The U.S. Role in Macroprojects," *Technology in Society* 5:51-68, 1983.
14. Tusso, B., Office of Economic Development, Paris, France, January 1988.
15. U.S. Congress, General Accounting Office, *Patent Policy: Recent Changes in Federal Law Considered Beneficial*, GAO/RCED-87-44 (Washington, DC: General Accounting Office, April 1987).
16. U.S. Congress, Office of Technology Assessment, *New Developments in Biotechnology, 4: U.S. Investment in Biotechnology* (Washington, DC: U.S. Government Printing Office, in press).
17. U.S. Congress, Office of Technology Assessment, *Intellectual Property Rights in an Age of Electronics and Information*, OTA-CIT-302 (Washington, DC: U.S. Government Printing Office, April 1986).
18. U.S. Congress, Office of Technology Assessment, *Research Funding As an Investment: Can We Measure the Returns?* OTA-TM-SET-36 (Washington, DC: U.S. Government Printing Office, April 1986).
19. U.S. Congress, Office of Technology Assessment, *Commercial Biotechnology: An International Analysis*, OTA-BA-218 (Springfield, VA: National Technical Information Service, January 1984).
20. Van Horn, C., personal communication, U.S. Patent and Trademark Office, Washington, DC, February 1988.
21. Wade, N., *The Nobel Duel* (Garden City, NY: Anchor/Doubleday, 1981).
22. Watson, J.D., *The Double Helix* (New York, NY: New American Library, 1969).
23. Yuan, R.T., International Trade Administration, U.S. Department of Commerce, *Biotechnology in Western Europe* (Washington, DC: U.S. Government Printing Office, 1987).

Appendixes

Appendix A

Topics of OTA Contract Reports

The following reports were prepared by outside contractors for the Office of Technology Assessment for this assessment. They are available on microfiche or as hard copy from the National Technical Information Service (NTIS), 5285 Port Royal Road, Springfield, VA 22161; tel: (703) 487-4650.

- **Mapping Our Genes Contractor Reports,** Vol. 1, Order No. PB 88-160 783/AS
  - —*Bibliometric Analysis of Work on Human Gene Mapping*, Samuel R. Reisher and Michael B. Albert, CHI Research/Computer Horizons, Inc.
  - —*Medical Implications of Extensive Physical and Sequence Characterization of the Human Genome*, Theodore Friedmann, Center for Molecular Genetics, School of Medicine, University of California, San Diego
  - —*Mapping the Human Genome: Some Ethical Implications*, Jonathan Glover, New College, Oxford University
  - —*Mapping and Sequencing the Human Genome: Considerations From the History of Particle Accelerators*, John L. Heilbron, University of California, Berkeley, and Daniel J. Kevles, California Institute of Technology
  - —*Mapping the Human Genome: Historical Background*, Horace Freeland Judson, The Johns Hopkins University
  - —*Long-Term Implications of Mapping and Sequencing the Human Genome: Ethical and Philosophical Implications*, Mark A. Lappé, College of Medicine, University of Illinois at Chicago
- **Mapping Our Genes Contractor Reports,** Vol. 2, Order No. PB 88-162 805/AS
  - —*The Mapping and Sequencing of Genomes: A Comparative Analysis of Methods, Benefits and Disbenefits*, Stephen M. Mount, Department of Biological Sciences, Columbia University
  - —*Mapping the Human Genome: Experimental Approaches for Cloning and Ordering DNA Fragments*, Richard M. Myers, Department of Physiology, University of California, San Francisco
  - —*Mapping and Sequencing the Human Genome in Europe*, Peter A. Newmark, *Nature*
  - —*Application of Human Genome Mapping for the Global Control of Genetic Disease*, Sir David J. Weatherall, Nuffield Department of Clinical Medicine, John Radcliffe Hospital, Oxford University
  - —*In Search of the "Ultimate Map" of the Human Genome: The Japanese Efforts*, Akihiro Yoshikawa, Berkeley Roundtable on the International Economy, University of California

OTA also sponsored two workshops during this assessment, and the transcripts of those workshops have been submitted to NTIS as:

- **Mapping Our Genes, Transcript of Workshop "Issues of Collaboration for Human Genome Projects,"** June 26, 1987, Order No. PB 88-162 797/AS; and
- **Mapping Our Genes, Transcript of Workshop "Costs of Human Genome Projects,"** Aug. 7, 1987, Order No. PB 88-162 813/AS.

The following report was written for OTA but was not sent to NTIS because it will be available to the public through a law journal article:

- "Sharing of Research Results in a Federally Sponsored Gene Mapping Project," Susan Rosenfeld, Science and the Law Committee, Association of the Bar of the City of New York, August 1987; to be published by the *Rutgers Computer and Technology Law Journal*, vol. 14, No. 2, 1988.

Appendix B

Estimated Costs of Human Genome Projects

Congress has primary responsibility for funding research through Federal agencies because of its responsibility for the national budget each year. Appropriating Federal funds for any special genome projects will therefore fall to Congress. These appropriations will express Congress' judgment regarding the relative value of genome projects. In setting appropriation levels, Congress will weigh the costs of the programs against their anticipated benefits (in economic and social terms) and will balance the value of proceeding against the costs of not doing so, as measured in lost benefits or opportunities.

Proposals for genome projects are intended to support research, but research needs are inherently unpredictable: Technological breakthroughs could dramatically diminish budget needs, and unanticipated obstacles could just as dramatically increase them. Estimates of near-term projects using existing technologies are necessarily more accurate than future projects that presume technological developments. Costs for some of the larger components, such as sequencing significant portions of human or nonhuman DNA, hinge on unit costs that are highly uncertain now and are rapidly changing due to technical advances (e.g., the cost of sequencing a single base pair of DNA). These uncertainties suggest that a 5-year budget plan is the best that can be produced, and projected costs for even the first 5 years might need to be substantially revised. The costs of human genome projects can be separated by components, although the boundaries between some of them are imprecise. The costs projected in this appendix are based on a process followed by OTA to generate estimates from internationally recognized experts.

OTA Cost Estimates

In order to better estimate potential costs of human genome projects, OTA held a workshop on August 7, 1987. At that workshop, there was apparent consensus on rough estimated costs of several components and confusion or disagreement about many others. A follow-up letter was sent to workshop participants and over 150 experts from executive agencies, universities, and corporations to confirm estimates made at the workshop and to expand them. Replies were received from over 70 persons. The revised cost estimates were externally reviewed by over 100 individuals and institutions in a draft report circulated in November 1987, and some minor revisions are based on comments received during this review. The resulting cost projections attempt to include most of the direct costs of research. They do not include indirect costs of university administration (although they do include administration in Federal agencies).

In some cases, it may prove possible to attract funding from the private sector—foundations, medical research institutes, or corporations. If so, Federal spending could be correspondingly reduced. In many cases, however, the Federal Government will eventually pay the full costs. If a company developed mapping and sequencing information or new instruments, for example, the first—and for a long time the predominant—users would remain researchers funded to do biomedical research by the Federal Government. This would be the case for most technologies developed as part of human genome projects (use by researchers being the primary goal of the enterprise). A company's investment would thus be charged back to the government by charging for use of information or purchase of instruments by the research community. In some cases there may be a market for products outside the biomedical research community. If so, the private sector funds could indeed displace government funds. Funding from research foundations, medical institutes, and other philanthropies would also, as a rule, substitute directly for government costs.

There was strong consensus about the importance and feasibility of improving the research infrastructure (databases and repositories) and generating genetic and physical maps of human and nonhuman chromosomes; there was substantial uncertainty about sequencing strategies and their associated costs. It is agreed that the need for new technologies is paramount, but there is disagreement about how much it would cost to develop them or how such efforts should be organized.

Discussion in the following sections reviews costs by component.

Computers and Computational Methods

Cost estimates for the necessary personnel, research, and equipment are $12 million per year for the early years, increasing to 15 percent of the overall budget as it exceeds $80 million annually. This would be *in* *addition to* continued support of existing databases and computer facilities. Spending should be relatively flat over time, because hardware will have to be purchased in early years and research will take an increasing proportion of the budget in later years. Hardware will have to be upgraded, however, so cost estimates are necessarily uncertain for future years. **While it is logical to link computational needs to human genome projects, funding devoted to storage of genetic data and sophisticated analysis of DNA will prove important in molecular biology even if maps are not completed and other human genome projects are not funded.**

Genetic Maps

Genetic mapping has been conducted for several years, and a rough map of human chromosomes already exists. Discussion at the OTA cost workshop centered on a map with two to four times the resolution of current maps. Subsequent letters and discussions have centered on a further increase in resolution, preferably such that a gene being studied would be separated from its closest DNA markers (on average) in only 1 of 100 family members. (Geneticists call this a 1-centimorgan map.) Estimates based on existing procedures yield annual costs of $6 million per year for 5 years. Since this is an existing technology and there are already facilities to do the mapping, a startup period is not needed. Funds saved from new methods could be devoted to automating the processes so further refinements of maps in humans and other organisms would be easier to construct in the future.

The two principal groups constructing human genetic marker maps to date have not been federally supported. One has used private corporate funds, and the other has been funded by the Howard Hughes Medical Institute (HHMI). HHMI-sponsored work is a nearly direct substitute for goverment funding. Future work would be of greater magnitude, however, and may require Federal investment. In the case of work supported by Collaborative Research (the largest corporate group) and other companies, the Federal Government will probably pay for access to the probes either as a lump sum (to obtain access for all federally funded researchers) or indirectly (as federally funded researchers pay for access to individual DNA markers or mapping services).

Physical Maps

Physical maps would be quite useful for future research. Ordered clone sets linked to them would be even more useful. Pilot projects on selected human chromosomes and on many lower organisms are in progress, and a useful set of ordered clones from all the human chromosomes may be feasible in the next 5 to 10 years.

Projections based on existing technology yield costs of $60 million for a usefully complete set of ordered cosmid clones over 5 years. New technologies may permit the creation of ordered sets of much larger DNA fragments (using yeast artificial chromosomes, YACs), and these would be extremely useful also. Costs of constructing ordered libraries composed of both cosmid and YAC clones are estimated at $70 million.

There are substantial uncertainties regarding both types of clones. Physical mapping of human chromosomes using cosmid clones has only begun in the last year, and therefore the rate and completeness of such mapping are highly uncertain. Mapping with yeast artificial chromosomes is much newer, although promising. The main uncertainty regarding YACs is not cost, but feasibility: If such mapping is possible, it would be substantially less expensive than mapping using cosmids (although cosmid maps might be needed for many research applications).

Ordered clone libraries are difficult to complete. Progress is rapid at first, but it is unlikely that a chromosomal region can be spanned without gaps between groups of continuous clones. Maps complete enough to be useful can be expected from several years' effort, but if truly complete maps are necessary, then efforts must be continued, perhaps at funding levels equal to those for initial construction. Half or more of the total effort may be required for the last 10 percent of the maps. Clone libraries with gaps are quite useful, however, because a chromosomal region of interest is likely to be represented even in incomplete libraries.

Cost estimates start at $10 million for the first year (building on current Federal expenditures), rise to $20 million in $5 million increments over 2 years, and then drop to $10 million (with the proviso that continued higher funding may be necessary if complete maps are deemed essential).

Projects To Link Genetic and Physical Maps

Identifying the parts of DNA that carry the instructions for making protein and integrating them into genetic and physical maps would be very useful. Which stretches of DNA are actually used to produce protein varies with the tissue (many genes are expressed differently in different tissues or stages of development). The likely process would be to make DNA copies (cDNA) of the RNA that is translated into protein from a variety of tissues (both healthy and diseased) and at various stages of development. Locating cDNAs on physical and genetic maps would result in cDNA maps.

Such maps could be used to pick out protein-coding regions along stretches of DNA of unknown function. This would make the physical map much more useful, by highlighting regions of particular interest, and would provide a missing step in the search for genes whose approximate location had been determined by genetic linkage maps. DNA sequencing might also begin by using cDNAs to select regions likely to be of interest (because they are known to produce protein). Maps of cDNA would give clues to a gene's function if the pattern of expression related to a known biological process. Comparison of cDNAs from human and other organisms can give clues to function by relating expression to degree of evolutionary relatedness. If a genetic disease is located in a certain chromosomal region and cDNA maps show that one DNA segment from that region is transcribed only in the tissue affected by the genetic disease, then the gene corresponding to the cDNA is a good candidate for the gene causing the disease. Maps of cDNAs have been suggested by several groups [2,3,11,12,13,15].

The first step in constructing cDNA maps would be to collect and organize existing sets of cDNA clones. New sets of cDNA clones could be made from missing tissues, disease states, developmental stages, or organisms. The various cDNAs could then be located on genetic linkage maps and physical maps. The cost of this process is highly uncertain, in part because the number of genes in human and many other organisms is not known. Those specifically asked about this component estimated that its costs would likely range from $2 million to $5 million per year, depending on how much work could be done by merely cataloging existing cDNA clone sets; how many new sets would have to be constructed; how many organisms, tissues, developmental stages, and disease states would be used as sources; and the extent of genetic and physical maps. The costs of cDNA mapping would increase with the increasing detail of genetic and physical maps. OTA estimates start from a base of $2 million, increasing annually by $1 million increments.

Resource Material Repositories

Estimated costs of storing the clone sets linked to physical maps, cell lines for genetic research, and the various DNA analytical materials for genetic mapping originally ran to over $250 million. The largest component, dwarfing all others, was the cost of storing the DNA clones linked to physical maps. Such storage costs are virtually prohibitive, and these estimates were dropped. Subsequent discussions with experts on storage of materials for molecular biology, specifically with persons at the American Type Culture Collection, yielded storage estimates an order of magnitude lower. The estimates summarized in table B-1 are for collection and storage of clone sets. Costs of dissemination would be borne by users through user fees. Costs of collecting and storing mutant cell lines and DNA analytical materials (such as probes) have not been included.

Sequencing

There is little consensus on how much DNA sequencing should be done as part of genome projects, particularly whether a complete human reference sequence should be an objective. **There is consensus, however, that sequencing technology is crucial and ripe for innovation**. Cost projections should become easier in 2 to 3 years, as the first automated DNA sequencing machines are improved; massive sequencing would not begin in most schemes for several years,

Table B-1.—OTA Budget Projections for Genome Projects (millions of dollars, adjusted to 1988)

Component	Year 1	Year 2	Year 3	Year 4	Year 5
Computers and analysis	12	12	17	24	29
Genetic maps	6	6	6	6	6
Physical maps	10	15	20	20	10[a]
cDNA maps	2	3	4	5	6
Resource material repositories	1	2	3	4	5
Sequencing	—	—	15[b]	30[b]	45[b]
Quality control	—	1	2	3	4
Technology development	10[b]	20[b]	50[b]	75[b]	100[b]
Training	4	6	8	10	12
Aministration	2	3	6	9	11
Total	47	68	131	186	228

[a] May require upward adjustment for map closure.

[b] Subject to considerable uncertainty, depending on technical improvements, strategy, and unit costs.

SOURCE: Office of Technology Assessment, 1988.

except as part of pilot projects and for DNA regions known to be of special interest [6]. The debate about sequencing involves disagreement about the costs of sequencing per base pair, the amount of redundancy necessary to make a sequence useful, the expected pace of technological improvements, which laboratory preparation steps are included, and how much DNA would be sequenced as part of human genome projects (rather than through traditional funding mechanisms). Estimates of the cost of sequencing vary widely, ranging from several pennies to several dollars per DNA base [6,16], but there is some agreement that costs would drop to $0.20 to $0.30 per base pair by the end of 1988, based on existing technologies. Some of the discrepancy in the estimates comes from including different components. The costs of special cloning procedures, preparing DNA, use of reagents, technician time, and capital costs of instrumentation should all be included in cost estimates.

Judgments about which technologies to use and how much sequencing should be performed are best made each year by an advisory committee with access to technical experts. Such judgments would presumably be based on costs, the availability of material to sequence, and consensus on which regions to sequence first. For OTA projections, a few assumptions have been made. For the first 2 years' budget, sequencing would be covered as technology development—performed on lower organisms or human chromosomal regions of known interest—for possible sequencing on a larger scale. For years 3 to 5, it would be based on sequencing one small chromosome per year at $0.20 per base pair ($30 million per year, based on threefold redundancy and 50 megabases per year), permitting a phase-in period for implementation of the technologies. This estimate is for purposes of budgeting only, however, and could prove wildly high or low. If the technologies for cloning, preparing DNA for sequencing, and finally determining a DNA sequence become significantly cheaper, as some experts predicted at the OTA workshop, the *amount* of DNA sequenced at that cost could be increased. If costs remain high, only a limited amount of DNA could be sequenced, according to priorities set by the oversight board. After the fifth year, the budget could go up or down in proportion to need.

Quality Control and Reference Standards

The large amounts of map and sequence information and new materials created by human genome projects will be useful only if the information is accurate and resource materials are cataloged reliably. If there are many different groups involved in the efforts, problems of quality control could impede useful applications. The scope and magnitude of this problem will become clear only when the technologies are defined and the results of mapping and sequencing efforts begin to accumulate. Special budget allocations for comparing results from different groups or to establish measurement standards may become necessary. Budget needs for quality control will be nil in the first year and will grow in early years until they constitute 5 percent of the overall budget. For initial estimates, it is projected to grow by $1 million per year from a base of zero.

Technology Development

Investments in methods and instruments associated with genome projects are likely to lead quickly to commercial applications. The objective for technology development is open-ended, however, and it could be either the largest component or a relatively small fraction of genome projects. Responses to OTA letters and drafts showed no consensus on the proper budget. Many scientists familiar with industrial development encouraged higher figures, while academic molecular biologists set lower ones. A maximum figure of $500 million to be spent over 5 years was mentioned at the OTA workshop, in line with recommendations of a committee established by the Department of Energy (DOE). There was some support for the alternative of devoting 25 percent of the total budget to technology development [6]. Minimum estimates were for a steady state of $20 million to $30 million. Several individuals noted the importance of developing technologies early on, while recognizing the need to keep early budgets realistically low because a new research program would require the accumulation of trained personnel and pilot work to provide a foundation for later work.

The approach used in OTA estimates is to increase funding from $10 million the first year to a stable figure of $100 million by yearly increments. Funding for biological instrumentation centers under the National Science Foundation might account for part of this, and methods or instruments of great interest to industry might lead to some cost sharing with private firms. If so, Federal funding could be reduced accordingly. Technology development funding, like sequencing, is among the most flexible of the proposed projects and could be adjusted by the oversight board and the congressional appropriations process.

Training of Personnel

Training of investigators and scientific exchange among participants are crucial and would include graduate and postgraduate fellowships, scientific workshops, and national scientific meetings. Some persons urge that fellowship funds be targeted to shortage areas, but others believe that targeted programs are less effective than untargeted ones for the best people in any relevant discipline. If training were targeted, it might include development of dual expertise in computers and molecular biology, organic chemistry and molecular biology, engineering and molecular biology, and clinical medicine and informatics or molecular biology. Training would also be needed for technicians, and for sabbaticals for scientists interested in shifting from their fields to genome projects. Workshops among participating groups and national symposiums to communicate results would permit rapid dissemination of new methods and insights. Exchange programs among industrial, national laboratory, and academic scientists would promote technology transfer. Training and personnel costs are estimated to merit 10 percent of each annual budget. For initial projections, funding might start at $4 million and increase yearly by $2 million.

Administrative Costs

Participants in the August 1987 OTA workshop estimated that 1 to 3 percent of each year's budget would be needed for administrative overhead. That estimate was subsequently increased to 5 percent in response to letters and after analyzing administrative costs at Federal research agencies. Administrative costs include operation of a national advisory board; oversight of databases, repositories, networks, and other services; setting instrumentation standards for cloning, mapping, and sequencing technologies; administration of grants and contracts; and other purposes. Some additional features would be unique to genome projects, for example, analysis of likely social impacts and ethical dilemmas created or intensified by genome projects. The need for such analysis has been explicitly noted in hearings and has been highlighted by research agency administrators and congressional staff. It could be obtained through grants to bioethicists, lawyers, economists, and social scientists for publications or workshops on various topics.

Summary

The costs of the components of human genome projects are projected in table B-1. These would start from a base of $47 million in fiscal year 1989 (if 1989 were the first year) and increase to $228 million in fiscal year 1993. It is not useful to project budgets beyond then, because technological development is so uncertain. The projected figures do not attempt to assign functions to particular agencies, merely to state overall direct research costs. Future budgets will need to be revised in light of actual appropriations.

History of Earlier Estimates

Perhaps the earliest evidence of a human genome project is found in a letter from Robert L. Sinsheimer, then Chancellor of the University of California, Santa Cruz (UCSC), to University of California President David Pierpont Gardner, on November 19, 1984. A potential benefactor had withdrawn support from a project, and Sinsheimer took the opportunity to provide a counterproposal that might interest the benefactor. In doing so, Sinsheimer suggested that a human genome institute be founded at UCSC, with startup costs of $25 million and an annual operating budget of $5 million. This was, in effect, the first cost estimate for a human genome project.

The letter from Sinsheimer to Gardner referred to an enclosure, later to be used as a basis for discussion at the May 1985 Santa Cruz Human Genome Workshop, in which the institute was formally proposed [7]. The proposal assumes existing mapping technologies and a continued rate of development of DNA sequencing speed equal to the exponential increase of the past decade. The proposal then concludes that "within a few years, the human genome could be reduced to an ordered set of cloned fragments" and that "50 technicians could approach completion of the [sequencing] project in 10-20 years." The proposal estimates the yearly support of each technician at $100,000, yielding an annual budget of $5 million and a total project cost of about $100 million. The proposal also calls for $25 million for startup facilities, and it distributes the operating money among the mapping and sequencing project itself (75 percent), developing techniques (10 percent), application to basic biology and medicine (10 percent), and education and training of students and other personnel (5 percent).

The Santa Cruz workshop displays similar optimism about the mapping aspect of a genome project, suggesting that a physical map could be completed by a 20-person group in 3 to 5 years [14]. The workshop also included discussion of a restriction fragment length polymorphism (RFLP), or genetic, map. This map could be achieved in "a few years" at a resolution finer than 50 centimorgans. Based on then current technologies, sequencing the 3 billion base pairs of the human genome was taken as "not feasible." The workshop went beyond the initial proposal and discussed details about the computer requirements for a project. There was, however, no explicit cost estimate for these details.

The next round of cost estimates came out of DOE's workshop in Sante Fe in March 1986. Appended to the workshop notes and the correspondence the workshop generated between the participants and DOE's Mark Bitensky was a cost estimate by Christian Burks of the Los Alamos National Laboratory. Burks calculates the person-years required for various aspects of the project, which, for a physical mapping and sequencing endeavor, including computer and administrative costs and assuming some sequencing advances, totals 3,505 person-years [5]. Allowing for hardware and overruns of his estimate, Burks concludes a genome project would cost between $0.5 and $2.5 billion.

The next major meeting, at Cold Spring Harbor, focused primarily on sequencing. The only estimate to issue from the discussion was the oft-quoted 30,000 person-years required to sequence the human genome one time through [8]. This estimate—translated into $3 billion by either $1 per base pair or $100,000 per person-year—was based solely on existing technology and was therefore obsolete within days of the conference, when the automated sequencer at the California Institute of Technology was announced [18].

By the middle of 1986, the Caltech sequenator had made it clear that advances in sequencing technology would drive costs down. HHMI's Informational Forum at the National Institutes of Health in Washington, D.C., continued to quote the 30,000 person-year estimate, but it also cautiously offered an estimate of 300 person-years, assuming a two-order-of-magnitude increase from automation [17]. The HHMI forum likewise gave a dual estimate for the physical map (200 person-years, or 30 to 40 with automation advances) [4], and for computer storage of sequence information ($0.30 per base pair, $0.03 with advances) [1].

Nine months later, DOE brought out its own cost estimates, presented as a yearly budget for a genome project. In the Health and Environmental Research Advisory Committee report, the subcommittee scientists estimate that sequencing, with redundancy for accuracy, would cost $60 million, assuming advances in automation [19]. Sufficient automation should be available 5 years hence [10]. The remainder of the budget is not described in detail, but it does specify that $500 million will go to various aspects of technological development, including mapping and informatics, assuming $100,000 per person-year of research [9]. The total for the DOE-proposed projects comes to $1.02 billion. Table B-2 presents a summary of estimates.

The National Research Council of the National Academy of Sciences established the Committee on Mapping and Sequencing the Human Genome, whose report was released in February 1988 [13]. That report represents the views of an exceptionally distinguished panel of experts from diverse scientific backgrounds. The panel members began their deliberations with widely differing knowledge of the state of gene mapping and divergent opinions about the merit of special research efforts. While writing the report, the panel reached a consensus that a special effort was merited and recommended additional funding of $200 million per year. This level would be reached over the initial 3 years. During the first few years, the budget would be roughly divided into $120 million for research in 10 or so multidisciplinary centers and numerous small research groups. Construction and materials would cost $55 million per year, and $25 million would operate repositories, databases, training, and administrative functions. In later years, the budget would increase for dedicated production of map and sequence data. This $200 million annual budget would continue until at least the year 2000.

Appendix B References

1. Bell, George I., comments at Informational Forum on the Human Genome, sponsored by Howard Hughes Medical Institute, July 23, 1986, at the National Institutes of Health, Bethesda, MD.
2. Berg, Paul, comments at Costs of Human Genome Projects, OTA workshop, Aug. 7, 1987.
3. Berg, P., Stanford University Medical Center, personal communication, January 1987.
4. Brenner, Sydney, comments at Informational Forum on the Human Genome, sponsored by Howard Hughes Medical Institute, July 23, 1986, at the National Institutes of Health, Bethesda, MD.
5. Burks, Christian, "The Cost of Sequencing the Complete Human Genome," Santa Fe Genome Sequencing Workshop, Mar. 3-4, 1986.
6. Costs of Human Genome Projects, OTA workshop, Aug. 7, 1987.
7. Edgar, Bob, Noller, Harry, and Ludwig, Bob, "Human Genome Institute: A Position Paper," enclosure in personal correspondence from Robert L. Sinsheimer, University of California at Santa Cruz, to David Pierpont Gardner (Nov. 19, 1984) distributed at the Santa Cruz Human Genome Workshop, May 24-25, 1985.
8. Gilbert, Walter, comments at Molecular Biology of *Homo sapiens*, Cold Spring Harbor Laboratories symposium, May 28-June 4, 1986.
9. Hood, L., California Institute of Technology, personal communication, June 1987.
10. Hood, L. California Institute of Technology, personal communication to David Guston, June 1987.
11. Karnei, R., and Stocker, J.T., Armed Forces Institute of Pathology, personal communication, December 1987.
12. McKusick, V.A., and Ruddle, F.H., "Toward a Complete Map of the Human Genome," *Genomics* 1:103-106, 1987.
13. National Research Council, National Academy of Sciences, *Report of the Committee on Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, 1988).

14. Notes and Conclusions, Santa Cruz Human Genome Workshop, sponsored by Department of Energy and Los Alamos National Laboratory, June 4, 1985.
15. Philipson, L., and Tooze, J., "The Human Genome Project," *BioFutur*, June 1987, pp. 94-101.
16. Roberts, L., "New Sequencers to Take on the Genome," *Science* 238:271-273, 1987.
17. Smith, David, remarks at Informational Forum on the Human Genome, sponsored by Howard Hughes Medical Institute, July 23, 1986, at the National Institutes of Health, Bethesda, MD.
18. Smith, Lloyd M., Sanders, Jane Z., Kaiser, Robert J., et al., "Flourescence Detection in Automated DNA Sequence Analysis," *Nature*, 321:674-678, 1986.
19. Subcommittee on the Human Genome, Health and Environmental Research Advisory Committee, *Report on the Human Genome Initiative*, prepared for the Office of Health and Environmental Research, Office of Energy Research, Department of Energy (Germantown, MD: DOE, April 1987).

Table B-2.—Comparison of Genome Cost Estimates (millions of dollars (M) or person-years (py))

Source	RFLP map	Physical map	Sequence	Computing	Repository	Other	Total
UCSC[a] position paper 11/19/84		"a few years"	500-1,000 py				$25 M facilities
UCSC workshop 5/24-5/85	"a few years" (<50 cM)	60-100 py	"not feasible"				
Sante Fe workshop 3/86		55 py	3,000 py	300 py + hardware		150 py administration	$500-2,500 M[a]
Cold Spring Harbor Symposium 5/28-6/2/86			30,000 py[b]				
HHMI/NIH Informational forum 7/23/86	200 py or 30-40 py[b]	30,000 py or 300 py[b]	$.30/bp or $.03/bp[b]				
DOE/HERAC 4/87			$60 M or 6,000 py[a]			$500 M technology	$1,020 M
NRC						$ 60 M/yr 10 centers[a] $ 60 M/yr grants and technology[a] development for small groups $ 55 M/yr[c] facility construction (early years, decreasing later) $ 25 M/yr administration, quality control, advisory committee functions $200 M/yr	$3,000 M (over 15 yrs.)
OTA[d] 8/87-1/88	$30 M (1cM)	$70 M YAC and cosmid	$60 M not complete	$12 M min 15% of total	$15 M	$10 M quality control $20 M linking projects 5% administration $255 M technology development $40 M training	$660 M (first 5 yrs. only)

[a] Assumes $100,000 per person-year.

[b] Not consensus figures but individual opinions.

[c] Money for facilities in early years would go to mapping and sequencing in later years.

[d] Estimates for first 5 years only. Does not assume complete reference sequence. For details, see text.

Abbreviations: DOE/HERAC—Health and Environmental Advisory Committee, Department of Energy; HHMI—Howard Hughes Medical Institute; NIH—National Institutes of Health; NRC—National Research Council; OTA—Office of Technology Assessment, U.S. Congress; UCSC—University of California at Santa Cruz.

SOURCES **UCSC:** personal communications from Robert Sinsheimer, Chancellor, UCSC, January 1987 and August 1987; Santa Cruz Human Genome Workshop (SCHGW), "Notes and Conclusions," June 4, 1985; and Bob Edgar, Harry Noller, and Bob Ludwig, "Human Genome Institute: A Position Paper," enclosure in personal correspondence from Robert L. Sinsheimer to David Pierpont Gardner (Nov. 19, 1984) and distributed for Santa Cruz Human Genome Workshop (May 24 to 25, 1985); **Sante Fe Workshop:** Christian Burks, "The Cost of Sequencing the Complete Human Genome," App. VI: Genome Sequencing Workshop, Sante Fe, NM, Mar. 3 and 4, 1986; **HERAC:** U.S. Department of Energy (DOE), *Report on the Human Genome Initiative*, Subcommittee on Human Genome of the Health and Environmental Research Advisory Committee, April 1987; **Cold Spring Harbor Symposium:** Walter Gilbert, comments at Cold Spring Harbor Laboratories Symposium, "Molecular Biology of *Homo sapiens*, May 28 to June 4, 1986; **HHMI/NIH symposium:** Remarks of George Bell, Sydney Brenner, and David Smith, at Howard Hughes Medical Institute Informational Forum on the Human Genome, July 23, 1986; **NRC:** National Research Council, National Academy of Science, Committee on Mapping and Sequencing the Human Genome, *Mapping and Sequencing the Human Genome* (Washington, DC: National Academy Press, Feb. 1988); **OTA:** "Costs of Human Genome Projects" Workshop, Aug. 7, 1987, with subsequent summary letter and review, October 1987, and external review of cost projections December 1987.

Appendix C

Participants in OTA Workshops

ISSUES OF COLLABORATION FOR HUMAN GENOME PROJECTS

Workshop Sponsored by the
Office of Technology Assessment
and the
Howard Hughes Medical Institute

June 26, 1987
2322 Rayburn House Office Building
United States Congress
Washington, DC

Chairman
William W. Lowrance, Ph.D.
Director
Life Sciences and Public Policy Program
The Rockefeller University

Bernadette Alford, Ph.D.
Legal and Regulatory Affairs
Collaborative Research

George F. Cahill, M.D.
Vice President for Scientific Training and Development
Howard Hughes Medical Institute

C. Thomas Caskey, M.D.
Director
Institute for Molecular Genetics
Baylor College of Medicine

James F. Childress, Ph.D.
Department of Religious Studies
University of Virginia

Susan Cozzens, Ph.D.
Department of Social Sciences
Illinois Institute of Technology

George Gould, Esq.
Associate Patent Counsel
Hoffman LaRoche

Geoffrey M. Karny, Esq.
Dickstein, Shapiro & Morin

Susan Rosenfeld, Esq.
Science and the Law Committee
Association of the Bar of the City of New York

Frank Ruddle, M.D.
Department of Biology
Yale University

Robert Smith, Ph.D.
History of Science Department
The Johns Hopkins University

Robert E. Stevenson, Ph.D.
Director
American Type Culture Collection

Charles E. Van Horn
Director
Organic Chemistry and Biotechnology
U.S. Department of Commerce, Group 120
Patent and Trademark Office

LeRoy Walters, Ph.D.
Director
Center for Bioethics
Kennedy Institute of Ethics
Georgetown University

83-630 0 - 88 - 7 QL 3

COSTS OF HUMAN GENOME PROJECTS

Workshop Sponsored by the
Office of Technology Assessment

August 7, 1987
Office of Technology Assessment Conference Center
Washington, DC

Chairman
Paul Berg, M.D.
Willson Professor
Department of Biochemistry
Stanford University Medical Center

Christian Burks, Ph.D.
Los Alamos National Laboratory

George F. Cahill, M.D.
Vice President for Scientific Training and Development
Howard Hughes Medical Institute

Anthony V. Carrano, Ph.D.
Biomedical Sciences Division
Lawrence Livermore National Laboratory

Helen Donis-Keller, Ph.D.
Collaborative Research, Inc.

Walter Gilbert, Ph.D.
Biological Laboratories
Harvard University

Leroy Hood, M.D.
Chairman
Division of Biology
California Institute of Technology

Keith McKenney, Ph.D.
Center for Chemical Physics
National Bureau of Standards

Mark L. Pearson, Ph.D.
Director of Molecular Biology
E.I. du Pont de Nemours & Co.

Robert E. Stevenson, Ph.D.
Director
American Type Culture Collection

John Sulston, D.Phil.
MRC Laboratory of Molecular Biology
Cambridge, England

James D. Watson, Ph.D.
Director
Cold Spring Harbor Laboratory

Appendix D

Databases, Repositories, and Informatics

Among the most useful products of genome projects will be information and materials—information about genes and their locations and sequences, and biological materials such as DNA fragments from chromosomes of known pedigree, ordered cosmids, and clones. Proper management of data and materials is essential to increase the efficiency and productivity of research and to reduce duplication of efforts so that genome projects can succeed in meeting the needs of medical scientists and molecular biologists in this century and the next.

Existing databases and repositories that gather, maintain, analyze, and distribute data and materials are already struggling to keep up with the exponential growth of molecular biology. Present capabilities will have to expand greatly to handle the increase of information resulting from a targeted set of genome projects. **While it is logical to link computational needs to genome projects, however, funding devoted to storage of genetic data and materials and to sophisticated analysis of DNA will prove important in molecular biology even if a major mapping and sequencing initiative is not undertaken**. Because the essential databases, repositories, and linking computer networks provide goods and services for the entire research community, the Federal Government has a long-standing tradition of supporting them and is in a unique position to further enhance the resources.

This appendix describes some existing databases and repositories and outlines present and future database needs relevant for human genome projects specifically and molecular biology in general.

Databases

Various databases exist that serve the needs of researchers in genome mapping and sequencing (see table D-1). One set of databases gathers, stores, and distributes information directly related to genetic maps and physical maps. Some databases specialize in map and sequence information from one specific genome—for example, there are databases exclusively devoted to the mouse, *E. coli* bacteria, drosophila, and nematode genomes—while others carry particular kinds of information from all the relevant genomes. Other databases gather data on the sequences and structures of proteins and amino acids that are not direct results of mapping and sequencing research but are necessary for addressing basic research problems underpinning genome research. **The data from the different types of maps and from different species have important interconnections, so it is essential that the information be linked for comparative studies.**

Genetic Maps

Genetic maps can be generated in several ways (see ch. 2). Pedigree analysis of linked traits yields a map in which traits can be ordered sequentially and with a rough estimate of the distance between them. RFLPs and other DNA probes can help link the traits with specific genes or regions of DNA to produce more refined maps. Maps of the functional regions within individual genes aid in the search for underlying causes of genetic diseases and for the mechanisms by which genes control development and function. Several different databases serve the different information needs for specific kinds of maps.

On-Line Mendelian Inheritance in Man (OMIM).—An atlas of human traits that are known to be inherited—expressed genes—has been compiled into a reference work known as *Mendelian Inheritance in Man*, which has been published in seven editions. The listing has been edited by Victor McKusick of The Johns Hopkins University since 1966. As of March 1, 1988, 4,336 traits had been identified as genetically based, including over 2,000 diseases.

Since 1986, the Howard Hughes Medical Institute (HHMI) has supported computerization of the list, and it is now accessible for on-line searches free of charge (4). It is cross-referenced in the Human Gene Mapping Library so that information on expressed genes can be linked to map data.

Human Gene Mapping Library (HGML).—Also called the New Haven Database, HGML consists of five linked databases—one each for map information, relevant literature, RFLP maps, DNA probes, and contacts (researchers with information on data or materials). In addition, the map database is linked to OMIM. All of the databases are cross-referenced, so that data about a gene or probe of interest can be drawn from all five during the same search (10).

DNA Nucleotide Sequences

Databases containing raw DNA sequences, information about the origin of the DNA segment sequenced

Table D-1.—Some Existing U.S. Databases and Repositories

	Location	Funding source	Annual budget[a]
Nucleotide sequence data:			
GenBank®	Los Alamos National Laboratory, Intelligenetics Corp., CA	NIH,[b] DOE, NSF, USDA	$3,500,000
Genetic map data:			
On-Line Mendelian Inheritance in Man (OMIM)	Johns Hopkins University Baltimore, MD	Johns Hopkins University, HHMI, NLM	$ 550,000[c]
Human Gene Mapping Library (HGML)	New Haven, CT	HHMI	$ 500,000
Protein and amino acid sequence and structure data:			
Protein Identification Resource (PIR)	National Biomedical Research Foundation Washington, DC	NIH[d]	$ 500,000
Protein Data Bank (PDB)	Brookhaven National Laboratory Upton, NY	NSF, NIH,[e] DOE	$ 260,000
Repositories:			
American Type Culture Collection (ATCC)/Human DNA Probe and Chromosome Library	Rockville, MD	NIH[f]	$ 300,000[e]
Human Genetic Mutant Cell Repository	Coriell Institute for Medical Research Camden, NJ	NIH[g]	$ 750,000

[a] Budget figures are approximate. Several of the databases have multiyear contracts; amount listed is the average yearly allotment.

[b] NIH sponsors of GenBank®, past and present, include the National Institute of General Medical Sciences (NIGMS), the Division of Research Resources (DRR), the National Institute for Allergy and Infectious Diseases, the National Cancer Institute, the National Library of Medicine, the National Eye Institute, and the National Institute of Diabetes and Diseases of the Kidney. The NIGMS administers the contract and coordinates the funding.

[c] The Johns Hopkins University contribution to OMIM is difficult to measure, because it includes many indirect factors (staff support, space, publication costs, etc.). HHMI contributes $318,000 and the NLM $100,000 annually.

[d] The NIH sponsor is DRR.

[e] NIH sponsors are NIGMS and DRR.

[f] NIH sponsors are the National Institute of Child Health and Human Development (NICHD) and DRR; DOE has contributed some funds through DRR.

[g] The NIH sponsor is NIGMS.

SOURCE: Office of Technology Assessment, 1988.

(which gene, which organism), and various annotations that summarize information about important features in the sequence (sites cut by DNA-cutting enzymes, regulatory sequences, protein-coding regions) will be directly affected by genome projects that emphasize sequencing. The major databases for nucleotide sequences are GenBank® and its European counterpart, EMBL (8). Each carries sequence data and related information for the human genome as well as bacterial, yeast, fruit fly, mouse, and other genomes. Since 1982, GenBank® and EMBL have split the task of data collection, with each database monitoring specific journals in molecular biology to locate and enter sequence data, and they cooperate closely in sharing and distributing it. They have recently been joined by the DNA Data Bank of Japan (DDBJ), which is in charge of monitoring Asian journals and contributing to the reciprocal exchanges. (DDBJ served primarily as an access node to GenBank® and EMBL starting in 1984, but did not start gathering its own data until 1987.)

GenBank®.—GenBank® originated at the DOE's Los Alamos National Laboratory in 1979 and started to receive funding from the NIH in 1982. It is the major U.S. database for nucleic acid sequence information from humans and other organisms (3). GenBank® is presently administered and receives a major portion of its funds from the National Institute of General Medical Sciences (NIGMS) of NIH. Data are entered and updated by curators at Los Alamos and are distributed by Intelligenetics Corp. (Mountain View, CA).

The amount of data contained in GenBank® has grown exponentially since its inception. In addition, the number of users has increased from a small set of one hundred or so who accessed it when the first NIH contract started to tens of thousands of scientists who now access either directly or through commercial distributors. GenBank®'s new 5-year contract, which took effect in October 1987, significantly increases funding to meet the growing demand.

Protein and Amino Acid Sequences and Structures

Databases that gather information on protein and amino acid structure and function are crucial for the application of genomics research to clinical and pharmaceutical problems, as well as for advancing the understanding of basic problems in biology—how genes function, how they code for proteins and enzymes, and how their protein products are structured and function (see ch. 2). The effects of map and sequence data on these databases will depend on the strategy followed for genome projects. For example, a concerted nucleotide sequencing effort would affect research on protein and amino acid structure more slowly than increased funding to researchers studying specific genes and their gene products—generally proteins (6).

Protein Identification Resource (PIR).—PIR is "a resource designed to aid the research community in the identification and interpretation of protein sequence information" (14). It contains sequence data for proteins and amino acids, with annotations that indicate known functional regions. PIR is run by the nonprofit National Biomedical Research Foundation and receives most of its funding from NIH's Division of Research Resources. Modest user fees cover the distribution costs; academic users pay a flat fee, while commercial users are charged by the amount of computer time they use. PIR has recently started cooperating with the Japan International Protein Database (JIPID) and the new European database, Martinsreid Institute for Protein Sequence Data (MIPS), to establish an international data network for protein sequences.

Protein Data Bank (PDB).—The Protein Data Bank was founded in 1971 as "an international computerized archive for structural data on biological macromolecules" (1). It gathers information on the atomic coordinates of the structure of nucleic acids, messenger RNA, amino acids, proteins, and carbohydrates that have been derived from crystallographic studies. Structural information is a vital link in the understanding of how proteins function, which eventually leads to knowledge of the mechanisms of genetic disease and suggests possible directions for rational drug design.

PDB is based at DOE's Brookhaven National Laboratory and supported primarily by NSF, with additional funds from the National Institute of General Medical Sciences of NIH. Modest user fees help cover the costs of distribution. Use of the database has been growing rapidly and is predicted to continue growing in parallel with human genome projects. Linking PDB with databases that contain genetic map and sequence information will enhance the long-term goals of human genome research (12).

Present and Future Needs

The many types of information that are produced in molecular biology necessitate the maintenance of a variety of specialized databases. At the same time, however, the information in different databases must often be combined in order to understand the full dimensions of any specific research problem. **It is crucial for the scientific community to be able to access information on a topic of interest from a variety of databases that may handle different aspects of the problem.** Thus databases must use standardized or easily translatable formats and they must be interconnected. The problem of format has been recognized and is being addressed in scientific meetings, by database advisers, and by funding agencies. Several programs are underway to improve the linkages between databases. An experimental project at the National Library of Medicine, discussed below, will develop a system to link a variety of databases relevant to molecular biology.

The speed with which data are entered into the databases has been a major concern. The exponential increase in data has not always been matched by increases in the support for databases and personnel to operate them, causing a lag time of several months or even years between the publication of data and their entry, in fully annotated form, into databases. If the lag time is excessive, the efficiencies of centralized data management and retrieval are lost. One solution that is being explored is the direct submission of data to the databases by the researchers as a requirement for publication in journals. At least one journal has already agreed to cooperate with GenBank® and EMBL in an attempt to speed acquisitions in this way (19). Another possibility is to encourage funding agencies to make the submission of data or materials to the appropriate databases a condition of receiving research grants. The automation of data entry will be necessary as the amount of data increases. Automated methods are already under development; the capacity to enter data may be built into some automated sequencing machines.

The timely exchange of data is also affected by issues of intellectual property rights and technology transfer. Open and rapid exchange of information and materials speeds research and is particularly important when the data have medical or clinical implications. If the data and materials become commercially valuable, however—and many researchers predict that they will—the values of open access and free exchange could clash with the desire to protect proprietary rights on potentially patentable data or materials. Because access to databases and repositories is international, there are concerns that U.S.-funded research could be commercialized by other countries. The problems are not intractable, however: There are several successful precedents of advance contracts that specify how data will be contributed to databases while protecting property rights (4,21). (See also ch. 8.)

A major problem faced by databases for the past decade has been insufficient funding for handling the exponential increase of data. Costs will continue to rise as more map and sequence data are generated. The government agencies and other organizations that support genome projects appear to recognize the importance of continued funding for relevant databases. For example, the increased budget in the new GenBank® contract (for 1987 through 1991) indicates that funding agencies are aware of the need to enhance database maintenance. An initiative within the National Library of Medicine to strengthen information resources for molecular biology and biotechnology (discussed below) should lend further support to databases needed for genome projects. The Howard Hughes Medical Institute has been particularly active in supporting database resources and networks to link them. It is essential that financial support continue to keep pace with the growing body of data.

Repositories

Genome projects will generate biological materials as well as sequence and map data. Access to these materials is a key element in making the map information useful. A scientist searching for a gene of unknown location would want to have access to a panel of DNA markers that could give an approximate location, then a more closely spaced set of markers to locate it more precisely. Once the gene's location was established on the genetic map, the investigator would select DNA clones covering that region of the human chromosomes from a repository, thus obtaining the DNA encoding the gene. Each of these steps would require access to a set of cloned DNA fragments. Existing repositories are hardly sufficient, but how much must be invested in them will depend on conclusions on the value of centralized sources rather than housing materials in individual labs.

Companies developing a new product derived from or related to a human gene would also wish to have access to such materials in many instances. Storage and handling of such DNA resources is thus a crucial function. The materials will be most widely useful if they are stored at national collection and storage facilities. DNA probes, vectors, and some other materials are best maintained at a facility such as the American Type Culture Collection (ATCC). Others, such as cell lines derived from individuals and families with genetic diseases, are stored in the Human Genetic Mutant Cell Repository in Camden, New Jersey. Other materials that are unlikely to have substantial demand from a wide variety of investigators might be stored at the laboratories that generated them and distributed on a more informal basis to those requesting them. Present methods and technologies for the amplification, characterization, storage, and distribution of materials are expensive and time-consuming; the costs of storage could become a major component of mapping and sequencing projects. Newer and cheaper storage methods will have to be developed as production of DNA fragments increases. The development of automated techniques for organizing, managing, and accessing materials will be necessary; research on automated repository management is already underway at ATCC and at DOE's Los Alamos National Laboratory (11, 21).

Even with the advent of automated repository management techniques, however, the high cost of storing and maintaining materials makes the selection of materials to collect particularly crucial. While it might be desirable to keep large collections of clones generated in an attempt to develop libraries of overlapping clones or contigs (see ch. 2), the curators of repositories and the scientists who use them will have to choose which materials are of utmost importance, and these decisions should be periodically reviewed (22,23).

American Type Culture Collection

The ATCC maintains a variety of different collections of animal, plant, and bacterial cell lines, hybridomas, phage, and recombinant DNA vectors, as well as an NIH-sponsored repository of human DNA probes and chromosome libraries (20). The collection of chromosome libraries includes materials from DOE's National Gene Mapping Library (see ch. 5). The ATCC amplifies and stores samples and distributes them, along with pertinent information, to investigators for a nominal fee. Investigators must agree not to use the materials for commercial purposes nor to sell them.

The repository maintains a database of information on the source and characteristics of the material in its collection. Its advisory committee has recommended that the database be included in a mapping database such as HGML.

Human Genetic Mutant Cell Repository

Sponsored by the National Institute of General Medical Sciences of NIH, the Human Genetic Mutant Cell Repository was founded in 1972 to maintain a collection of well-characterized human cell cultures (2,17). The cultures are available to investigators worldwide at a nominal fee. The repository contains over 4,000 individual cultures, which represent more than 400 genetic diseases and 700 to 800 chromosomal aberrations (7). The curators of the collection have increasingly sought to include material from multigenerational family groups for linkage analysis; the repository now maintains cell lines from the Venezuelan Huntington's pedigree (see box 7-A) and others such as cystic fibrosis families, families with fragile X-linked mental retardation, and so on.

Data Analysis, Informatics, and Computer Resources

Development of analysis methods to search for and compare sequence information, to predict sequences that code for proteins and the structures of those proteins, and to aid in other aspects of the analysis of data from genome projects will eventually need to utilize parallel processing techniques and the capacity of supercomputers. Most researchers agree that the hardware to tackle the complex problems of sequence analysis and comparison already exists but that satisfactory software must be developed. The DOE, the NIH, the NLM, and the NSF support various programs and grants for the development of software to represent and analyze data and for the development of computer resources such as supercomputing centers and computer networks. Several of these resources are described below. Numerous private firms are developing or marketing computer programs that search databases or analyze data on nucleic acid or protein sequences.

BIONET™

BIONET™ is a nonprofit computer network run by Intelligenetics, Inc. (Mountain View, CA) and funded by the Division of Research Resources of NIH and by modest user fees (13). Its goals are to "provide computation assistance in data analysis and problem solving to molecular biologists and researchers in related field, to serve as a focus for the development and sharing of new software, and to promote rapid sharing of information and collaboration among a national community of scientists" (9). BIONET™ provides access to several major databases (GenBank®, EMBL, PIR, PDB, and databases of restriction enzymes and plasmid vectors) as well as to software for analyzing nucleic acid and protein sequences. The network also aids communication between its members through a series of bulletin boards on topics of user interest and through an electronic mail system. BIONET™ serves users in the United States, Canada, and Europe.

National Biotechnology Information Center

The National Biotechnology Information Center is an initiative to develop and enhance a range of tools for molecular biology information that is being sponsored by the National Library of Medicine (NLM) (18). The project is presently the subject of several authorizations bills but has already received some appropriations for a range of projects, including the building and maintenance of databases, developing a comprehensive listing of existing databases, and improving information retrieval systems. NLM has already developed a prototype of a retrieval system, called the Information Retrieval Experiment (IRX) that connects data from several different databases and graphic and visual sources. For example, a database search for a specific disease gene will yield information on whether the gene has been mapped, the map of the gene in graphic form, bibliographic information on publications about the map, as well as information on clinical symptoms, diagnosis, and visual representations of affected patients (X-rays, diagrams, photos, and so on). The NLM initiative will enhance the management of data from genome projects and will forge links between information from many areas of molecular biology to aid in basic and biomedical research (15). The NLM is in an advantageous position to coordinate database activities through its expertise in handling information through existing literature databases such as MEDLINE.

The Matrix of Biological Knowledge Workshop

The Matrix of Biological Knowledge Workshop, a month-long conference held during the summer of 1987, was an attempt to formulate models and make recommendations for the organization of knowledge and data from all disciplines in biology (16). It was sponsored by the NIH, the DOE, the Sloan Foundation, and the Santa Fe Institute.

The workshop grew out of the efforts of a committee sponsored by the NIH that attempted to set forth and evaluate models used in biomedical research. Several scientific meetings prior to the workshop had addressed the particular complexities of biologicial data; at the workshop, biologists, computer scientists, and database experts actually tried to work out some of the problems raised at earlier meetings. Participants at the workshop issued the following general recommendations:

> . . . that support for a centrally coordinated effort to establish a knowledge base of databases in the biological sciences be aggressively pursued; that the current independent efforts to establish inter-database structures and analysis tools be coordinated with a long-term view towards maximum integration; . . . that these coordinated efforts incorporate the most up-to-date computer science and analytical methods; and finally, that these activities directly involve the experimental and biotechnology communities in order to ensure the utility of the ensuing developments (16).

These recommendations appear to reinforce the direction of ongoing efforts in agencies that sponsor databases. The specific recommendations issued by working groups in each of seven broad categories may prove useful for the future management of databases in all of biology.

Appendix D References

1. Abola, E.E., Bernstein, F.C., and Koetzle, T.F., "The Protein Data Bank," in P.S. Glaeser (ed.), *The Role of Data in Scientific Progress* (New York, NY: Elsevier Science Publishers, 1985).
2. Aronson, M.M., Miller, R.C., Nichols, W.W., et al., "Breakpoint Map of Human Translocation Cell Cultures Available From the NIGMS Human Genetic Mutant Cell Repository," *Journal of Cytogenetics and Cell Genetics* 30:179-189, 1981.
3. Burks, C., Fickett, J.W., Goad, W.B., et al., "The GenBank® Nucleic Acid Sequence Database," *Computers Applications in Bioscience* 1:225-233, 1985.
4. Cahill, G.C., Vice President for Scientific Training and Development, Howard Hughes Medical Institute, Bethesda, MD, personal communication, December 1987.
5. Cassatt, J., National Institutes of Health, personal communication, April 1987.
6. George, D., Protein Identification Resource, National Biomedical Research Foundation, Washington, DC, personal communication, December 1987.
7. Greene, A.E., Human Genetic Mutant Cell Repository, Coriell Institute for Medical Research, Camden, NJ, personal communication, March 1988.
8. Hamm, G., and Cameron, G., "The EMBL Data Library," *Nucleic Acids Research* 14:5-9, 1986.
9. *Introduction to BIONET™* Mountain View, CA: Intelligenetics, Inc., 1987).
10. Kidd, K., Lecture at "Genomics Meeting: Assessing the Repository, Informatics, and Quality Control Needs," Lawrence Livermore National Laboratory, Livermore, CA, Aug. 26-27, 1987.
11. Knobeloch, D., and Beugelsdijk, T.J., "Automated Sample Management: Organizing, Managing, and Accessing the DNA Fragments Generated in the Process of Mapping and Sequencing the Human Genome," in *Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects*, see ref. 23.
12. Koetzle, T., Protein Data Bank, Brookhaven National Laboratory, Upton, New York, NY, personal communication, September 1987.
13. Kristofferson, D., "The BIONET™ Electronic Network," *Nature* 325:555-556, 1987.
14. Ledley & Barker, Protein Identification Resource project description, November 1987.
15. Masys, D., National Library of Medicine, Bethesda, MD, personal communication, December 1987.
16. Morowitz, H.J., and Smith, T. (eds.), *Report of the Matrix of Biological Knowledge Workshop* (Santa Fe, NM: Santa Fe Institute, 1987).
17. "The National Institute of General Medical Sciences Human Genetic Mutant Cell Repository," *Somatic Cell and Molecular Genetics* 12:421, 1986.
18. National Library of Medicine, "Biotechnology Information: A Plan for the National Library of Medicine," unpublished report, 1987.
19. "A New System for Direct Submission of Data to the Nucleotide Sequence Data Banks," *Nucleic Acids Research* 15, No. 18, 1987.
20. Nierman, W.C., Benade, L.E., and Maglott, D.R., *American Type Culture Collection: NIH Repository of Human DNA Probes and Libraries* (Rockville, MD: American Type Culture Collection, 1987).
21. Stevenson, R., American Type Culture Collection, Rockville, MD, personal communication, October 1987.
22. U.S. Congress, Office of Technology Assessment, "Costs of Human Genome Projects," workshop held Aug. 7, 1987 (see app. A).
23. U.S. Department of Energy, Office of Health and Environmental Research, and National Institutes of Health, *Repository, Data Management, and Quality Assurance Needs for the National Gene Library and Genome Ordering Projects*, unpublished report from a workshop, August 1987.

Appendix E

Bibliometric Analysis of Human Genome Research

Computer Horizons, Inc. (CHI) was hired by the Office of Technology Assessment to conduct a bibliometric analysis of work on human gene mapping, including an international bibliography of the most relevant literature. This bibliography centered on an examination of the growth of relevant scientific literature keyed to the words "Gene or Genes or Genetic," "Marker or Linkage or Map," and "Human." Additional key word combinations included "Human Chromosome," "Human DNA Sequence," "Human Nucleic Acid Sequence," "Human Restriction Fragment Length Polymorphism," and various combinations designed to select papers on methods and techniques of DNA analysis.

The use of publication counts as a measure of research activity is part of the field of bibliometrics. A growing body of research has demonstrated the usefulness of bibliometric techniques: Counts of scientific papers and the numbers of citations to them have been shown to be indicators of research productivity. Limitations to bibliometric techniques do exist, particularly in balancing the treatment of non-English publications. Since the literature of science is dominated by English-speaking researchers, there is an inherent bias against citations of foreign-language publications. In the present case, however, the primary purpose of the literature search was to provide indicators of growth rather than to develop specific bibliographies. As such, the analysis clearly demonstrated a rapid growth in the scientific literature related to mapping and sequencing the human genome, and an acceleration of this growth over very recent years.

Over 11,000 entries of relevant literature were presented in the bibliography, which scanned appropriate publications from 1977 through 1986. The literature search included journals published in English, French, German, Dutch, Italian, Polish, Japanese, Spanish, Russian, Bulgarian, Swedish, Finnish, Norwegian, Danish, and Hebrew. All entries were subsequently grouped by OTA into the country or region of origin to identify national and regional trends in research. The regions included the United States, Western European countries, Japan, and other non-European countries. The table below presents the results. The data were used as the basis for figures 7-1 and 7-2 and table 7-1 in chapter 7, which display the growth in the total number of articles on human gene mapping and sequencing and the breakdown by country or region.

Annual Publications in Human Genetics:
Articles Published on Human Genes or Genetic Markers and Linkage Maps

Year	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986
United States	187	218	235	314	308	364	487	577	689	818
Japan	7	11	17	22	32	45	39	58	67	85
Western Europe										
Denmark	6	14	9	7	12	7	9	12	21	25
Federal Republic of Germany	20	23	14	33	42	41	51	69	78	100
Finland	3	6	8	7	7	5	9	12	15	14
France	21	34	36	42	59	57	64	70	94	114
Italy	8	8	9	15	24	15	44	39	44	66
Netherlands	15	25	17	20	13	18	25	25	50	45
United Kingdom	32	49	57	46	66	88	97	126	184	185
Other	31	19	46	30	45	44	62	69	70	92
Other Non-European countries										
Australia	2	8	11	17	18	22	24	23	20	38
Canada	12	17	17	28	14	29	26	38	60	68
Eastern Europe and U.S.S.R.	23	17	21	38	36	33	36	51	60	62
South Africa	0	6	7	8	6	3	4	6	16	9
Other	20	20	35	33	33	41	32	42	87	63
Uncertain	32	41	79	75	57	87	61	81	95	101
Total	419	516	618	735	772	899	1,070	1,298	1,650	1,885

SOURCE: Office of Technology Assessment, 1988.

Appendix F

Acknowledgements

OTA wishes to acknowledge the many people contributed to the prepartion of this report. Members of the advisory panel and contractors are listed at the front of the report. Workshop participants are listed in appendix C. Others helped by commenting on drafts of the report, providing information or advice, or consenting to interviews with OTA staff. OTA particularly thanks those at the National Institutes of Health, the Department of Energy, the National Science Foundation, the Howard Hughes Medical Institute, and other organizations for their invaluable efforts to inform OTA about their activities. Our gratitude is extended to:

Carlos Abella
Embassy of Spain

Bruce M. Alberts
University of California at San Francisco

Duane Alexander
National Institute of Child Health and Human Development

Norman Anderson
Large Scale Biology

Wyatt Anderson
University of Georgia

David G. Baldwin
Tetrarch Inc.

David Baltimore
Whitehead Institute

Phil Beachy
Carnegie Institution of Washington

George I. Bell
Los Alamos National Laboratory

Celeste Berg
Carnegie Institution of Washington

Fred Bergmann
National Institute of General Medical Sciences

Michel Bernon
Embassy of France

Ralph Bledsoe
Domestic Policy Council

Lars Bolund
University of Aarhus, Denmark

Judith Bostock
Office of Management and Budget

David Botstein
Genentech, Inc.

James E. Bowman
University of Chicago

Elbert Branscomb
Lawrence Livermore National Laboratory

Douglas Brutlag
Stanford University

Martin Buechi
Embassy of Switzerland

John Burris
National Academy of Sciences

Andrew Bush
U.S. Senate

Mark Cantley
Commission of the European Economic Community

Charles R. Cantor
College of Physicians and Surgeons of Columbia University

C. Thomas Caskey
Baylor College of Medicine

James Cassatt
National Institute of General Medical Sciences

James W. Chamberlin
U.S. Embassy, Brazil

Andrew T. L. Chen
Centers for Disease Control

James F. Childress
University of Virginia

George M. Church
Harvard University Medical School

Mary Clutter
National Science Foundation

Stanley Cohen
Stanford Medical School

Francis Collins
University of Michigan

P. Michael Conneally
Indiana University Medical Center

Cheryl Corsaro
National Institutes of Health

Alan Coulson
MRC Molecular Biology Laboratory
Cambridge, United Kingdom

Charles L. Coulter
Division of Research Resources
National Institutes of Health

David Cox
University of California at San Francisco

James F. Crow
University of Wisconsin

Terry Curtin
U.S. Senate

Kay E. Davies
Oxford University

Bernard D. Davis
Harvard University Medical School

Ronald Davis
Stanford University

Michael Dean
National Cancer Institute

Larry L. Deaven
Los Alamos National Laboratory

Albert de la Chapelle
University of Helsinki

Enrique Martin del Campo
Organization of American States

Charles DeLisi
Mount Sinai School of Medicine

Vincent DeVita
National Cancer Institute

Denis Dewez
Embassy of Belgium

Russell F. Doolittle
University of California, San Diego

Janet Dorigan
Office of Science and Technology Policy

Renato Dulbecco
The Salk Institute

Irene Eckstrand
National Institute of General Medical Sciences

Peter Farnham
American Society for Biochemistry and Molecular Biology

James Fickett
Los Alamos National Laboratory

John C. Fletcher
University of Virginia

Donald S. Fredrickson
National Institutes of Health

Jean Frezal
Hospital for Sick Children, Paris

F. Edwin Froehlich
U.S. Senate

Robert Fujimura
Oak Ridge National Laboratory

David George
National Biomedical Research Foundation

Frank Gibson
The Australian National University

Paul Gilman
U.S. Senate

Alan Goldhammer
Industrial Biotechnology Association

George M. Gould
Hoffmann-La Roche, Inc.

Denise Greenlaw
U.S. Senate

Santiago Grisolia
University of Kansas Medical Center

Ralph Hardy
Boyce Thompson Institute

Wendy Harris
Johns Hopkins University Press

Nemat Hashem
Ain Shams University
Cairo, Egypt

Cyril G. Hide
South African Embassy

C. Edgar Hildebrand
Los Alamos National Laboratory

Joseph R. Hlubucek
Embassy of Australia

Sverker Hogberg
Swedish Embassy

Michael Hunkapiller
Applied Biosystems, Inc.

Thomas Isenhour
Utah State University

Trefor Jenkins
University of the Witwatersrand
Johannesburg, South Africa

Chalmers Johnson
University of California, Berkeley

Irving S. Johnson
Eli Lilly & Co.

Eric T. Juengst
The Pennsylvania State University

Robert F. Karnei
Armed Forces Institute of Pathology

Steven Keith
U.S. Senate

Michael J. Kelly
Intelligenetics, Inc.

Kenneth Kemphues
Cornell University

Thomas Koetzle
Protein Data Bank, Brookhaven National Laboratory

George S. Kopp
U.S. House of Representatives

Arthur Kornberg
Stanford University

David Kristofferson
BIONET/Intelligenetics

Louis Kunkel
Howard Hughes Medical Institute, Boston

Alphonse Lafontaine
Ministerie van Volksgezondheid en van het Gezin
Brussels, Belgium

Roe Laird
Ministry of State/Science and Technology
Canada

Eric Lander
Whitehead Institute

Norman Latker
U.S. Department of Commerce

Eileen Lee
U.S. House of Representatives

Hans Lehrach
Imperial Cancer Research Fund

Rachel Levinson
National Institutes of Health

Jack G. Lewis
University of Southern California

Jiayao Li
The Embassy of the People's Republic of China
Washington, D.C.

Donald A.B. Lindberg
National Library of Medicine

John Logsdon
George Washington University

Jeffrey T. Lutz
U.S. Embassy, Indonesia

Jerold R. Mande
U.S. Senate

Emmanuele Mannarino
U.S. Embassy, Italy

Daniel R. Masys
National Library of Medicine

Kenichi Matsubara
Osaka University
Japan

Sunil Maulik
BIONET/Intelligenetics

George Mazuzan
National Science Foundation

Jack McConnell
Johnson & Johnson

Victor A. McKusick
The Johns Hopkins University

Mortimer L. Mendelsohn
Lawrence Livermore National Laboratory

Bruce Merrifield
U.S. Department of Commerce

Bradie Metheny
Tricom, Inc.

Jerome P. Miksche
U.S. Department of Agriculture

Sankar Mitra
Oak Ridge National Laboratory

Jan Mohr
Institute of Clinical Genetics, University of Copenhagen

Robert G. Morris
U.S. Embassy, Argentina

Diane Morton
Cornell University

Jay Moskowitz
National Institutes of Health

Tom Murray
Case Western Reserve University

DeLill Nasser
National Science Foundation

Dorothy Nelkin
New York University

Norrine Noonan
Office of Management and Budget

Stephen O'Brien
National Cancer Institute

Maynard V. Olson
Washington University School of Medicine

Gilbert S. Omenn
University of Washington

Stuart Orkin
Howard Hughes Medical Institute, Boston

Joseph Osterman
U.S. Army Medical Research and Development
Command

Joseph Palca
Nature

M. Iqbal Parker
University of Cape Town

Jane Peterson
National Institute of General Medical Sciences

Ulf Petterson
University of Uppsala, Sweden

Kate Phillips
Council on Governmental Relations

Betty Pickett
Division of Research Resources
National Institutes of Health

Maya Pines
Howard Hughes Medical Institute

George Poste
Smith Kline & French Laboratories

Michael Probert
Medical Research Council
United Kingdom

Theodore T. Puck
Eleanor Roosevelt Institute for Cancer Research

Robert Rabin
National Science Foundation

Alan S. Rabson
National Cancer Institute

Lisa J. Raines
Industrial Biotechnology Association

William F. Raub
National Institutes of Health

Jeremy Rifkin
Foundation for Economic Trends

Jerry H. Roberts
National Institutes of Health

Leslie Roberts
Science

Richard Roberts
Cold Spring Harbor Laboratories

Thomas Rollins
U.S. Senate

Leon E. Rosenberg
Yale University

Luigi Rossi-Bernardo
Consiglio Nazionale della Ricerche, Italy

John Roth
University of Utah

Lesley Russell
U.S. House of Representatives

Joseph Sambrook
University of Texas Health Science Center

Mona Sarfaty
U.S. Senate

D. Schmitz
U.S. Embassy, Federal Republic of Germany

David Schwartz
Carnegie Institution of Washington

Charles Scriver
McGill University and
Science Council of Canada

Gerald Selzer
National Science Foundation

Leroy C. Simpkins
U.S. Embassy, Mexico

Michael Simpson
Library of Congress

Robert Sinsheimer
California Institute of Technology

Mark Skolnick
University of Utah

Bent Skou
Royal Danish Embassy

David A. Smith
U.S. Department of Energy

Lloyd Smith
University of Wisconsin

Robert Smith
The Johns Hopkins University

Temple Smith
Dana Farber Cancer Institute

Rand Snell
U.S. Senate

M. Anne Spence
University of California, Los Angeles

J. Claiborne Stephens
Human Gene Mapping Library

Robert E. Stevenson
American Type Culture Collection

Irene Stith-Coleman
Library of Congress

J. Thomas Stocker
U.S. Department of Defense

Gary Stormo
University of Colorado

William Szkrybalo
Pharmaceutical Manufacturers'
Association

Bruna Tesso
Organization for Economic Cooperation and Development

Ignacio Tinoco
University of California, Berkeley

Kevin M. Ulmer
SeQ, Ltd.

Victor L. Urquidi
El Colegio de Mexico, A.C.

Paul Van Belkom
National Health and Medical Research Council
Commonwealth of Australia

Dorle Vawter
University of Minnesota Health Center

Robert Walgate
The Panos Institute

Bertil Wennergren
Swedish Commission on Genetic Engineering

Norman Whiteley
Applied Biosystems, Inc.

Robert Williamson
Saint Mary's Hospital, London

JoAnn Wise
University of Illinois

Carl Woese
University of Illinois

John Wooley
National Science Foundation

James B. Wyngaarden
National Institutes of Health

Douglas J. Yarrow
Embassy of the United Kingdom

Philip Youderian
University of Southern California

Debbie Zucker
U.S. Senate

Appendix G

Glossary

Alleles: Alternative forms of a genetic locus; alleles are inherited separately from each parent (e.g., at a locus for eye color there might be alleles resulting in blue or brown eyes).

Amino acid: Any of a group of 20 molecules that combine to form proteins in living things. The sequence of amino acids in a protein is determined by the genetic code.

Autoradiography: A technique that uses X-ray film to visualize radioactively labeled molecules or fragments of molecules; used in analyzing length and number of DNA fragments after they are separated by gel electrophoresis.

Autosome: A chromosome not involved in sex determination. The diploid human genome consists of 46 chromosomes, 22 pairs of autosomes and 1 pair of sex chromosomes.

Base pair: Two nucleotides (adenosine and thymidine or guanosine and cytidine) held together by weak bonds. Two strands of DNA are held together in the shape of a double helix by the bonds between base pairs.

Centimorgan: A unit of measure of recombination frequency. One centimorgan is equal to a 1 percent chance that a genetic locus will be separated from a marker due to recombination in a single generation. In human beings, 1 centimorgan is equivalent, on average, to 1 million base pairs.

Cloning: The process of asexually producing a group of cells (clones), all genetically identical to the original ancestor. In recombinant DNA technology, the process of using a variety of DNA manipulation procedures to produce multiple copies of a single gene or segment of DNA.

Complementary DNA, cDNA: DNA that is synthesized from a messenger RNA template; the single-strand form is often used as a probe in physical mapping.

Contigs: Groups of clones representing overlapping, or contiguous, regions of a genome.

Crossing over: The breaking during meiosis of one maternal and one paternal chromosome, the exchanging of corresponding sections of DNA, and the rejoining of the chromosomes.

C-value paradox: The lack of correlation between the amount of DNA in a haploid genome and the biological complexity of the organism. (C-value refers to haploid genome size.)

Determinism: The theory that for every action taken there are causal mechanisms such that no other action was possible.

Diploid: A full set of genetic material (two paired sets of chromosomes), one from each parental set. All cells except sperm and egg cells have a diploid set of chromosomes. The diploid human genome has 46 chromosomes. Compare *haploid*.

DNA, deoxyribonucleic acid: The molecule that encodes genetic information. DNA is a double-stranded molecule held together by weak bonds between base pairs of nucleotides. There are four nucleotides in DNA: adenosine (A), guanosine (G), cytidine (C), and thymidine (T). In nature, base pairs form only between A and T and between G and C, thus the sequence of each single strand can be deduced from that of its partner.

DNA probes: Segments of single-strand DNA that are labeled with a radioactive or other chemical marker and used to identify complementary sequences of DNA by hybridizing with them. See *hybridization*.

DNA sequence: The relative order of base pairs, whether in a stretch of DNA, a gene, a chromosome, or an entire genome.

Domain: A discrete portion of a protein with its own function. The combination of domains in a single protein determines its unique overall function.

Double helix: The shape in which two linear strands of DNA are bonded together.

Electrophoresis: A method of separating large molecules (such as DNA fragments or proteins) from a mixture of similar molecules. An electric current is passed through a medium containing the mixture, and each kind of molecule travels through the medium at a different rate, depending on its electrical charge and size. Separation is based on these differences.

Enzyme: A protein that acts as a catalyst, speeding the rate at which a biochemical reaction proceeds but not altering its direction or nature.

Eukaryote: Cell or organism with membrane-bound, structurally discrete nucleus and other well-developed subcellular compartments. Eukaryotes include all organisms except viruses, bacteria, and blue-green algae. Compare *prokaryote*.

Eugenics: Attempts to improve hereditary qualities through selective breeding. See *positive eugenics*, *negative eugenics*, *eugenics of normalcy*.

Eugenics of normalcy: Policies and programs intended to ensure that each individual has at least a minimum number of normal genes.

Exons: The protein-coding DNA sequences of a gene. Compare *introns*.

Gamete: Mature male or female reproductive cell with a haploid set of chromosomes (23); that is, a sperm or ovum.

Gene: The fundamental physical and functional unit of heredity. A gene is an ordered sequence of nucleotides located in a particular position on a particular chromosome. See *gene expression*.

Gene expression: The process by which a gene's blueprint is converted into the structures present and operating in the cell. Expressed genes include those that are transcribed into mRNA and then translated into protein and those that are transcribed into RNA but not translated into protein (e.g., transfer and ribosomal RNAs).

Gene families: Groups of closely related genes that make similar products.

Gene product: The biochemical material, either RNA or protein, made by a gene. The amount of gene product is used to measure how active a gene is; abnormal amounts can be correlated with disease-causing genes.

Genetic code: The sequence of nucleotides, coded in triplets along the mRNA, that determines the sequence of amino acids in protein synthesis. The DNA sequence of a gene can be used to predict the mRNA sequence, and the genetic code can in turn be used to predict the amino acid sequence.

Genetic engineering technologies: See *recombinant DNA technologies*.

Genetic linkage map: A map of the relative positions of genetic loci on a chromosome, determined on the basis of how often the loci are inherited together. Distance is measured in centimorgans.

Genetics: The study of the patterns of inheritance of specific traits.

Genome: All the genetic material in the chromosomes of a particular organism; its size is generally given as its total number of base pairs.

Genome projects: Research and technology development efforts aimed at mapping and sequencing some or all of the genome of human beings and other organisms.

Genomic library: A collection of clones made from a set of overlapping DNA fragments representing the entire genome of an organism. Compare *library*.

Haploid: A single set of chromosomes (half the full set of genetic material), present in the egg and sperm cells of animals and in the pollen cells of plants. Human beings have 23 chromosomes in their reproductive cells. Compare *diploid*.

Homeo box: A short stretch of nucleotides whose sequence is virtually identical in all the genes that contain it. It has been found in many organisms, from fruit flies to human beings. It appears to determine when particular groups of genes are expressed in the development of the fruit fly.

Human gene therapy: Insertion of normal DNA directly into cells to correct a genetic defect.

Human Genome Initiative: Collective name for several projects begun in 1986 by DOE to 1) create an ordered set of DNA segments from known chromosomal locations, 2) develop new computational methods for analyzing genetic map and DNA sequence data, and 3) develop new techniques and instruments for detecting and analyzing DNA.

Hybridization: The process of joining two complementary strands of DNA, or of DNA and RNA, together to form a double-stranded molecule.

Informatics: The study of the application of computer and statistical techniques to the management of information. In genome projects, informatics includes the development of methods to search databases quickly, to analyze DNA sequence information, and to predict protein sequence and structure from DNA sequence data.

International technology transfer: Movement of inventions and technical know-how across national borders.

Introns: The DNA sequences interrupting the protein-coding sequences of a gene that are transcribed into mRNA but are cut out of the message before it is translated into protein. Compare *exons*.

Karyotype: A photomicrograph of an individual's chromosomes arranged in a standard format showing the number, size, and shape of each chromosome; used in low-resolution physical mapping to correlate gross chromosomal abnormalities with the characteristics of specific diseases.

Library: A collection of clones in no obvious order whose relationship can be established by physical mapping. Compare *genomic library*.

Linkage: The proximity of two or more markers (e.g., genes, RFLP markers) on a chromosome; the closer together the markers are, the lower the probability that they will be separated during meiosis and hence the greater the probability that they will be inherited together.

Locus: The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position. Some restrict use of *locus* to regions of DNA that are expressed. See *gene expression*.

Marker: An identifiable physical location on a chromosome (e.g., restriction enzyme cutting site, gene, RFLP marker) whose inheritance can be monitored. Markers can be expressed regions of DNA (genes) or some segment of DNA with no known coding function but whose pattern of inheritance can be determined.

Meiosis: The process of two consecutive cell divisions in the diploid progenitors of sex cells. Meiosis results in four rather than two daughter cells, each with a haploid set of chromosomes.

Messenger RNA, mRNA: A class of RNA produced by transcribing the DNA sequence of a gene. The mRNA molecule carries messages specific to each of the 20 amino acids. Its role in protein synthesis is to transmit instructions from DNA sequences (in the nucleus of the cell) to the ribosomes (in the cytoplasm of the cell).

Multifactorial or multigenic disorders: See *polygenic disorders*.

Mutation: Any change in DNA sequence that results in a new characteristic that can be inherited. Compare *polymorphism*.

Negative eugenics: Policies and programs intended to reduce the occurrence of genetically determined disease.

Nucleotide: A subunit of DNA or RNA consisting of a nitrogenous base (adenine, guanine, thymine, or cytosine in DNA; adenine, guanine, uracil, or cytosine in RNA), a phosphate molecule, and a sugar molecule (deoxyribose in DNA and ribose in RNA). Thousands of nucleotides are linked to form the DNA or RNA molecule. See *DNA*, *base pair*, *RNA*.

Oncogene: A gene, one or more forms of which is associated with cancer. Many oncogenes are involved, directly or indirectly, in controlling the rate of cell growth.

Physical map: A map of the locations of identifiable landmarks on DNA (e.g., restriction enzyme cutting sites, genes, RFLP markers), regardless of inheritance. Distance is measured in base pairs. For the human genome, the lowest-resolution physical map is the banding patterns of the 24 different chromosomes; the highest-resolution map would be the complete nucleotide sequence of the chromosomes.

Polygenic disorders: Genetic disorders resulting from the combined action of alleles of more than one gene (e.g., heart disease, diabetes, and some cancers). Although such disorders are inherited, they depend on the simultaneous presence of several alleles, thus the hereditary patterns are usually more complex than those of single-gene disorders. Compare *single-gene disorders*.

Polymorphism: Difference in DNA sequence among individuals. Genetic variations occurring in more than 1 percent of a population would be considered useful polymorphisms for genetic linkage analysis. Compare *mutation*.

Positive eugenics: The achievement of systematic or planned genetic changes to improve individuals or their offspring.

Prokaryote: Cell or organism lacking membrane-bound, structurally discrete nucleus and subcellular compartments. Bacteria are examples. Compare *eukaryote*.

Protein: A large molecule composed of chains of smaller molecules (amino acids) in a specific sequence; the sequence is determined by the sequence of nucleotides in the gene coding for the protein. Proteins are required for the structure, function, and regulation of the body's cells, tissues, and organs, and each protein has a unique function. Examples are hormones, enzymes, and antibodies.

Recombinant DNA technologies: Procedures used to join together DNA segments in a cell-free system (an environment outside of a cell or organism). A recombinant DNA molecule can enter a cell and replicate there, either autonomously or after it has become integrated into a cellular chromosome.

Replication: The synthesis of new DNA strands from existing DNA. In human beings and other eukaryotes, replication occurs in the nucleus of the cell.

Resolution: Degree of molecular detail or a physical map of DNA, ranging from low to high.

Restriction enzyme, endonuclease: A protein that recognizes specific, short nucleotide sequences and cuts DNA at those sites. There are over 400 such enzymes in bacteria that recognize over 100 different DNA sequences. See *restriction enzyme cutting site*.

Restriction enzyme cutting site: A specific nucleotide sequence of DNA at which a restriction enzyme cuts the DNA. Some sites occur frequently in DNA (e.g., every several hundred base pairs), others much less frequently (e.g., every 10,000 base pairs).

RFLP, restriction fragment length polymorphism: Variation in DNA fragment sizes cut by restriction enzymes; polymorphic sequences that are responsible for RFLPs are used as markers on genetic linkage maps.

Ribosomal RNA, rRNA: A class of RNA found in the ribosomes of cells.

RNA, ribonucleic acid: A chemical found in the nucleus and cytoplasm of cells; it plays an important role in protein synthesis and other chemical activities of the cell. The structure of RNA is similar to that of DNA. There are several classes of RNA molecules, including messenger RNA, transfer RNA, ribosomal RNA, and other small RNAs, each serving a different purpose.

Sex chromosomes: The X and Y chromosomes in human beings that determine the sex of an individual. Females have two X chromosomes in diploid cells; males have an X and a Y chromosome.

Single-gene disorders: Hereditary disorders caused by a single gene (e.g., Duchenne muscular dystrophy, retinoblastoma, sickle cell disease). Compare *polygenic disorders*.

Somatic cells: Any cells in the body except reproductive cells and their precursors.

Technology transfer: The process of converting scientific knowledge into useful products.

Transcription: The synthesis of mRNA from a sequence of DNA (a gene); the first step in gene expression. Compare *translation*.

Transfer RNA, tRNA: A class of RNA having structures with triplet nucleotide sequences that are complementary to the triplet nucleotide coding sequences of mRNA. The role of tRNAs in protein synthesis is to bond with amino acids and tranfer them to the ribosomes, where proteins are synthesized according to the instructions carried by mRNA.

Translation: The process in which the genetic code carried by mRNA directs the synthesis of proteins from amino acids. Compare *transcription*.

Vector: DNA molecule originating from a virus, a bacterium, or the cell of a higher organism used to carry additional DNA base pairs; vectors introduce foreign DNA into host cells, where it can be reproduced in large quantities. Examples are plasmids, cosmids, and yeast artificial chromosomes.

Index

Index

Acid Precipitation Task Force, 120
acquired immune deficiency syndrome, 64, 72, 96, 117
adenosine, 21-22
adenosine deaminase, 61
agriculture, genome mapping implications for, 73
alanine, codon, 23
Alberts, Bruce, 107
albinism, 149
albumin, 61
alcoholism, 86
aldolase, chromosome assignment of, 33
Alexander, Duane, 94 n.1
alleles, 27, 28, 58
alpha globin, 61
alpha interferon. 64
Alzheimer's disease, 65, 95, 146
American Society for Biochemistry and Molecular Biology, 127
American Type Culture Collection, 101, 141, 190, 192
amino acids
  codon, 23
  databases, 97, 98, 190-191
  generation of, 22-23
anemias, 64, 136
antibodies, 21
apoliproprotein E, 61
Applied Biosystems, Inc., automated DNA sequencer, 47, 108, 138
applied research, government controls on, 87
arginine, codon, 23
Argonne National Laboratory, 122
Armed Forces Institute of Pathology, 103
arthritis, 64
Ashburner, Michael, 42
Asian nations, interest in genome projects, 8
asparagine, codon, 23
aspartic acid, codon, 23
atrial natiuretic factor, 64
Australia, genome research, 8, 133, 148, 195
automation of DNA sequencing
  DOE initiative for, 7-8
  by Japan, 9, 47-48, 137-138
  new technologies, 46-48
  robotic devices, 48, 137-138
  standard setting for equipment for, 103
autoradiography/autoradiographs
  automated scanning of, 48
  use in DNA sequencing, 46
  use in physical mapping, 32, 40, 45
  use in RFLP mapping, 29
autosomes
  karyotyping of, 32
  mapping of genetic loci on, 27, 30, 33-34

bacteria
  DNA sequencing of, 45
  E. coli, 41, 45, 47, 100
  genome mapping, 4, 40, 41
  haploid DNA content, 25
  mitochondria similarities to, 71
  S. typhimurium, 45
bacteriophage T4, genomic map of, 40
Baltimore, David, 124
basic research
  government restriction of, 87
  value of, 81
Baylor College of Medicine, 97
bears, 36, 69
behavior, human, genetic factors in, 85-86
Berg, Paul, 126
beta globin, 61, 65, 70
beta interferon, 64
Billings, John Shaw, 97
Bio-Rad Laboratories, automated DNA sequencer, 47-48
biomedical research
  HHMI support for, 8
  NIH funding, 6, 95
Biomolecular Engineering Programme, 140
BIONET™, 193
biotechnology
  databases, 97
  European research programs in, 140-141
  international competitiveness in, 11
  NBS support for, 103
Biotechnology Action Program, 140-141
Biotechnology Research and Innovation for Development and Growth in Europe, 140-141
bison, 72
blood pressure disorders, 64
Brenner, Sydney, 147
burn treatment, 64
Botstein, David, 126
Bush, Vannevar, 102

C-value paradox, 24-25
Caenorhabditis elegans
  amount sequenced, 47
  genome mapping, 42
  genome size, 47
California Biotechnology, probe development, 58
California Institute of Technology, automated DNA sequencer, 47, 102
Cambridge University, genome mapping project, 42
Canada, genome research, 8, 133, 148, 195
cancer
  mutation-induced, 25
  polygenic nature, 62
  treatment, therapeutic agents, 64
catalase, 61
cataract surgery, 64
cDNA
  clones, 59-63, 67
  libraries, 59
  mapping, 30, 32, 44, 63
  restriction enzyme cutting, 35
cell culture, *see* somatic cell hybridization

Cell Line Two-Dimensional Gel Electrophoresis Database, 97
cell receptors, 21-22
Center for the Study of Human Polymorphism
  collaborative efforts of, 8, 106, 143-144, 146, 149
  family pedigree data set, 58, 143, 146
  funding for, 106
  mission, 145
Centers for Disease Control, 103
chicken, haploid DNA content, 25
Chiles, Lawton, 97
cholesterol, low-density lipoprotein, 58
chromosome markers
  for single-gene diseases, 57
  funding for research on, 6
  mapping through family inheritance patterns, 27
  maps, low-resolution, 4
  use for genetic linkage studies, 6, 27, 28, 44
chromosomes
  banding patterns, 30, 33, 34-35, 40, 42, 45, 56
  crossing over (recombination), 26, 27
  deletion of, 25, 31, 32
  diploid number, 21
  Drosophila melanogaster salivary gland, 30, 32
  duplication, 25
  E. coli, 41
  gene assignment to, 31
  haploid number, 21
  hybrids, single, 31
  inversion, 26
  isolation techniques, 31-32, 43
  of clinical significance, 44
  phage lambda, 36, 39, 42, 56, 100
  polytene, 42
  sorting, 31-32, 37, 47, 56, 97, 100
  species similarities in, 34-35, 36
  translocation, 25-26, 31, 32
  yeast artificial, 36-37, 39, 43, 56
chromosomes, human
  1, 27
  4, 28
  6, 148
  7, 31, 44, 62
  9, 33, 148
  10, 32, 33
  11, 32
  13, 34
  16, 31, 44, 100, 148
  17, 31, 33
  19, 31, 44, 100
  21, 35, 44, 95, 100, 145-146
  22, 44, 145
  average size, 37
  number, 3
  resemblance to primate chromosomes, 34-35
  X, 24, 27, 30, 31, 44, 63, 100, 148
  Y, 24, 30, 31, 145
chronic granulomatous disease, 57, 59, 62
Church, George, 44-45
cloning/clones
  access to and ownership of, 147
  automation of, 47, 48
  banding patterns, 40
  cDNA, 59-63
  disease-associated genes, 57, 59, 60
  of DNA fragments, 39, 72
  drug development through, 62-63
  E. coli, 41
  fingerprinting method for ordering, 42
  fruit fly chromosomes, 42
  gene isolation by, 31, 59
  libraries of, 39, 42, 59, 62-63; *see also* contig maps
  microdissection, 42
  NIH grants for, 94
  ordering of, 38-40, 41, 42
  overlapping, 39, 42, 62, 100
  phage lambda chromosomes, 36, 39, 42, 56, 100
  repositories, 97, 101, 115
  S. cerevisiae, 42
  vectors, 35-37, 38-39, 42, 56, 67, 100
  yeast artificial, 36-37, 39, 43, 56, 157
Cold Spring Harbor Laboratories Conference, 6
collaboration on genome research
  by Australia, 148
  by Center for the Study of Human Polymorphism, 8, 106, 143, 146, 149, 157
  center-based vs. networking, 156
  databases and repositories, 8, 139, 158-159
  DOE, 157
  existing frameworks, 155-157
  European, 142, 150
  International Human Gene Mapping Workshops, 29, 157
  international journals, 157-158
  organizational options, 152-155
  precedents for international scientific programs, 150-152
  views on, 152-153
  Washington University-RIKEN, 157
Collaborative Research, Inc.
  DNA probe development, 58, 108
  RFLP linkage map, 6, 30
collagen, 61, 67
color-blindness, 21
Columbia University
  mapping of *E. coli* genome, 41
  mapping of human chromosome *21*, 44, 100
Compton, Arthur Holly, 99
computers, computational methods, and software
  artificial intelligence, 96
  costs, 180-181
  for DNA sequencing, 65, 97
  gene mapping applications to, 57, 146
  networking, 156
  NIH funding for improvements in, 95, 96
  see also databases; informatics
Concertation Unit for Biotechnology in Europe, 140
consortia
  of Federal/private interests, authority for, 16, 121
  funding, 122
  goals, 121
  intellectual property rights, 122

  Midwest Plant Biotechnology Consortium, 122
  national, to administer genome projects, 14-15, 121-123
  peer review, 122
  two-tiered system, 122
contig mapping/maps
  construction, 39, 40
  correlation with large-fragment restriction maps, 42
  forward genetics applications, 61
  nematode, 42, 43
  reverse genetics applications, 62
  strategies, 43-44
  yeast, 42
controversial issues
  Big Science vs. small science, 125-128
  DNA sequencing, extent of, 4, 6, 44, 57, 79, 81
  feasibility of genome mapping, 4
  quotes on, 126
  resolution of genome mapping, 3, 79, 81, 88
  see also ethical issues
corn, haploid DNA content, 25
Coulson, Alan, 44-45, 147-148
Crick, Francis H.C., 3, 21
cysteine, codon, 23
cystic fibrosis, 57, 58, 62, 149
cytidine, 21-22

Dana Farber Cancer Institute, 97
databases
  access to and ownership of, 12, 82, 102, 128, 134, 139, 146; *see also* technology transfer
  Cell Line Two-Dimensional Gel Electrophoresis, 97
  CODATA Hybridoma Databank, 141
  DNA Data Bank of Japan, 139, 158
  DNA fingerprints, 80
  European support of, 141
  funding for, 7, 12, 96-97, 141, 190
  Genatlas, 157
  GenBank®, 46, 96, 98, 109, 115, 139, 142, 154, 158, 190
  genetic maps, 24, 98, 106, 189-190
  government protection of, 87
  HHMI, 7, 8, 98, 106
  Human Gene Mapping Library, 106, 189-190
  importance, 4, 9
  international collaboration on, 8, 139, 158-159
  Japan Protein Information Database, 159
  linking of, 98
  management of, 12
  Martinsreid Institute for Protein Sequence data, 159
  MEDLARS/MEDLINE, 97
  mouse genetics, 106
  National Library of Medicine, 7, 8, 12, 96, 97
  needs for, 191-192
  nucleotide sequence data, 46, 96, 98, 141, 158, 190
  On-Line Mendelian Inheritance in Man, 24, 98, 106, 189-190
  Protein Data Bank, 190-191
  Protein Identification Resource, 97, 98, 158-159, 190-191
Dausset, Jean, 145-146
DeLisi, Charles, 100, 153
Denmark, national genome research efforts, 133, 143, 195
deoxyribonucleic acid, *see* DNA listings
Department of Defense, biomedical research resources, 104
Department of Energy
  funding for genome projects, 7, 96, 100
  Health and Environmental Research Advisory Committee report, 101-102
  interest in massive sequencing, 9
  international research collaboration, 157
  as lead agency for genome projects, 12, 14, 116, 117
  mission, 7, 99-100
  Office of Health and Environmental Research, 99-101
  organization, 99, 117-118
  peer review, 101, 118
  recommendations for genome projects, 4, 11, 100
  research supported by, 100, 117
  workshops sponsored by, 6, 100
  see also Human Genome Initiative
determinism, effect of genome mapping on, 86
development, *see* human physiology and development
diabetes, 62
diseases
  infectious, 64
  linking mapping and sequencing data to, 104
  see also genetic diseases; and specific diseases
DNA
  amount relative to organism complexity, 24-25
  C-value paradox, 24-25
  cloning in plasmids, 36-37, 39
  complementary, *see* cDNA
  discovery, 3
  electrophoretic separation of, 37-39
  expendable fraction, 25, 57
  fingerprints, 89; *see also* genetic screening
  fragmentation of, 37-39
  mitochondrial, 71
  oldest human samples, 72
  polymerase, 45
  recombinant technology, *see* recombinant DNA technology
  replication process, 21-22
  structure, 3, 21-22
  transcription to mRNA, 23-24
  see also chromosomes
DNA markers, *see* chromosome markers
DNA probes
  automated synthesis of, 47, 48
  cDNA, 28-29, 32, 33, 59-61, 63
  companies developing, 58
  fluorescently labeled, 46-48
  for genetic disease diagnosis, 58-59
  in *in situ* hybridization, 33
  number needed to complete human linkage map, 29
  oligonucleotides, 48
  radioactively labeled, 28-29, 32-33, 40, 44-46, 58
  reliability, 58
  for RFLP markers, 28-29, 56, 58, 61-62

  synthetic, 48, 59-60
  use to clone genes, 60
DNA Segment Library, 97
DNA sequence/sequencing
  automation of, 47-48
  commercialization, 82-83, 133, 138-139
  computer-assisted, 65
  controversies, 4, 6, 44, 79; *see also* ethical issues
  costs, 6, 182-183
  database, 46, 96, 98, 190
  definition, 3, 21
  directly from genomic DNA, 45
  E. coli, 41, 100
  enhanced fluorescence detection method, 46, 47
  exons, 59, 61, 63, 65, 69
  expenditures, federal, 8
  facilities for, 13
  government role in, 87
  homeo box, 67-68
  importance, 9
  introns, 25, 30, 61, 65, 69-70
  longest stretch determined, 46
  of mitochondria, 71
  multiplex, 44-46
  mutation detection applications, 56
  NIH funding for, 8
  rate, 46
  repeated, 25, 28, 43, 57
  RFLP mapping required for, 37-39
  scanning tunneling microscopy for, 46
  selective amplification without prior cloning, 45-46
  species comparisons, 68-70
  steps, 47
  strategies, 44-45
  technologies, 44-47
  variations, 28, 29
  VNTR, 29
Domestic Policy Council, 8, 105, 109, 119
Donis-Keller, Helen, 30
Down's syndrome, 32, 35, 58, 95, 146
Drosophila melanogaster
  amount sequenced, 47
  genome mapping, 42-43
  genome size, 47
  salivary gland chromosomes, 30, 33
drugs and pharmaceuticals, development, 62-63
Duchenne muscular dystrophy, 57-59, 61-63
Duffy blood group, 27
Dulbecco, Renatto, 100, 126, 145
dwarfism, 64
dystrophin, 63

EG&G Biomolecular, automated DNA sequencer, 47
E.I. du Pont de Nemours & Co., automated DNA sequencer, 47
electrophoresis, *see* gel electrophoresis
England, *see* United Kingdom
enzymes
  functions, 22
  see also specific enzymes
epidermal growth factor, 64
Epstein-Barr virus, 46
erythropoietin, 64
Escherichia coli
  amount sequenced, 47
  genome mapping, 41, 43, 100
  genome size, 47
ethical issues
  academic freedom, 87
  access to and ownership of databases and repositories, 16, 82, 88
  access to and use of genetic information, 79-80
  attitudes and perceptions of ourselves and others, 85-86
  commercialization, 16, 82-83, 133
  diagnostic/therapeutic gap, 83
  eugenics, 81, 84-85, 88, 143-144
  genetic fingerprinting, 80
  government role in mapping and sequencing, 87, 88
  international competitiveness, 87-88, 133
  physician practice, 83
  reproductive choices, 83-84, 88
  responsibility for considering, 123-124
eugenics
  negative, 85, 143-144
  of normalcy, 85
  positive, 84-85
eukaryotes, 70-71
Europe, Eastern, interest in genome projects, 8, 133, 143, 195
Europe, Western, genome sequencing and mapping activities, 139-148; *see also* specific countries and organizations
European Economic Community, genome research, 139-141, 153
European Molecular Biology Laboratory, 8, 139, 141-142, 158
European Molecular Biology Organization, 8, 141
European Research Coordination Agency, 142, 145-146
European Science Foundation, 142, 156
evolution, *see* molecular evolution

facilities for genome research
  bioprocess engineering, 102
  data handling, European needs, 142
  DOE funding for, 100
  flow cytometry, 32, 97; *see also* specific national laboratories
  need for, 10, 128
  NSF biology centers, 8, 102-103, 109
factor IX, 61
factor VII, 61
factor VIII:C, 64
familial hypercholesterolemia, 56, 58, 149
family pedigree projects
  CEPH data set on, 58, 136, 146
  Danish, 143
  Egyptian, 134, 136
  on mental illness, 156
  South African, 149

  use in genetic linkage mapping, 27, 33, 58, 61
  Venezuelan, Huntington's disease, 63-64, 134-136, 143, 146
fatalism, 86
Federal Advisory Committee Act, 124
Federal Republic of Germany, genetics research, 8, 133, 143-144, 195
fibroblast growth factor, 64
Finland, national genome research effort, 133, 144
flow cytometry
  enhanced fluorescence detection in, for DNA sequencing, 46
  extraction of whole chromosomes by, 37
  facility, 32
France
  Center for the Study of Human Polymorphism, 33, 58, 144-146
  genome projects, 8, 144-145
  published genome research, 133, 195
fruit fly
  developmental regulation in, 67
  Drosophila melanogaster, 30, 33, 42-43, 47
  genome mapping, 42-43
  haploid DNA content, 25
  human DNA sequences compared with, 68
  lethal mutations in larval stage, 42
funding for genome projects
  advisory body for determining, 124
  of consortia, 122
  databases, 7, 12, 96-97, 190
  determinants, 98
  determinants of congressional appropriations, 11-12
  DNA marker studies, 6
  DOE, 7, 96, 100, 118, 190
  European Economic Community, 139-143
  HHMI, 7, 190
  international, 8
  NIH, 6, 7, 94-98, 117, 155, 190
  NSF, 7, 8, 96, 190
  pluralism in, 13, 15, 119
  priority setting, 10
  private vs. federal, 79, 83
  recommendations, 4, 11-12, 107
  through a lead agency, effects of, 12-13
  through a national consortium, 14
  USDA, 190
  West German, 144

Gall, Joseph, 126
Galton, Francis, 84
gamma interferon, 64
gel electrophoresis
  database, 97
  DNA separation for physical mapping, 37, 39, 45
  polyacrylamide, 45
  pulsed-field, 37, 44, 56
  in RFLP mapping, 28, 37, 58
GenBank®, 46, 96, 98, 109, 115, 139, 142, 154, 158, 190
gene expression
  control of, 57
  steps in, 23-24
  study centers, 144
gene products
  functions of, 67
  with potential as therapeutic agents, 62, 64
gene therapy, 64, 141
genes
  biochemical identification, 62
  in a chromosome band, number, 33
  color-blindness, 21
  definition, 3, 21, 24
  dosage mapping, 34
  encoding ribosomal RNAs, detection, 33
  expressed, 24, 30
  families of, 25, 70
  functions, approaches to understanding, 66-67, 73
  homeotic, 68
  isolation techniques, 31, 33, 59-62
  largest, 63
  linked, 26, 34; *see also* genetic linkage mapping/maps
  mapping, *see* genetic linkage maps
  species similarities in, 34
  structure/function relationships, study of, 144
  see also human genes
genes, human
  aldolase, 33
  chromosomal locations known, 4
  number of loci identified, 24, 30
  number per haploid genome, 24
  sizes, 61
genetic code
  definition, 21--24
  for amino acids, 23
genetic diseases
  chromosomal locations of genes for, 4
  clinical services for, 100
  companies developing DNA probes for diagnosis of, 58
  correlating gross chromosomal abnormalities with, 32
  diagnostic information, physician handling of, 83
  family pedigree studies, 61, 63
  HHMI support of research on, 8
  isolation of genes associated with, 59-62
  mechanisms, 4
  not associated with biochemical defects, 61
  polygenic, 62
  RFLP markers for, 28, 56, 58
  single-gene, 57, 88
  see also specific diseases
genetic information
  access to and use of, 79-80, 82, 84
  causes of changes in, 25-26
  insurer use of, 81, 83
  organization and function, 21-26
genetic linkage mapping/maps
  autoradiography use in, 29
  autosomes, 27
  costs, 181-182
  databases, 24, 98, 106, 189-190
  disease diagnosis applications, 56, 58, 62
  distance measurements on, 27

  early attempts, 4, 6, 21
  electrophoretic technology in, 28
  family pedigree data in, 27, 33, 58, 61
  HHMI funding for, 7
  medical applications, 56, 58, 62-64
  number of markers needed to complete, 29-30
  projects to link physical maps with, 181-182
  purpose, 26-27
  recombinant DNA technology use in, 28
  resolution, 62
  reverse genetics applications, 62
  of RFLP, 28-30, 62
  somatic cell hybridization for, 27
  X chromosome, 27
genetic locus, *see* chromosome marker
genetic screening
  ethical questions about, 80, 88
  for missing children, 80
  for proof of paternity, 80
genetic selection, *see* eugenics
genetics
  definition, 21
  forward, 59-61, 62-63
  HHMI funding for, 7
  molecular, NIH research resources activities related to, 97
  NIH funding for, 95
  population, 72-73
  reverse, 59, 61-62
Genetics Institute, robotic devices for DNA sequencing, 48, 108
Genome Corp., physical mapping project, 108
genome mapping
  agricultural applications, 73
  application in developmental studies, 42, 65
  automation, 47
  determinism and, 86
  distance measurements in, 40
  evolutionary applications, 68-72
  facilities, 13
  importance, 9
  international efforts and cooperation, 8, 9, 150-159; *see also* specific countries
  resolution levels in, 56, 79
  see also genetic linkage mapping/maps; physical mapping
genome mapping, human
  commercialization, 82-83, 138-139
  controversies, 3, 4, 6, 9, 44, 55, 57, 102
  government role in, 87
  priorities for, 88
  scale of efforts, 5, 24
  strategies, 43-46
genome mapping, nonhuman
  bacteria, 4, 40, 41, 44
  fruit fly, 42-43, 44
  importance, 9, 44, 107
  international efforts, 8, 42
  nematodes, 4, 42, 44
  plants, 73, 136, 149
  yeast, 4, 41-42, 44
genome projects
  accountability to Congress, 13, 14, 124
  administration of, 12-15, 115-123, 184
  advisory board structure for, 123-124
  appropriations for, *see* funding
  benefits, 11, 55, 56, 133, 172-174
  Big Science v. small science approach, 120, 125, 127-128
  center-based vs. networking, 156
  collaboration on, 150-159, *see* also collaboration on genome research
  commercialization potential, 82-83, 133, 138-139, 151, 165
  common features, 7
  component nature, 4, 6, 10
  congressional oversight, 15-17
  congressional role in, 11-17
  consortium structure for, 14-15, 121-122
  cooperation among agencies, 9, 15, 118-119
  costs, 11-12, 4, 47, 180-185
  definition, 4
  displacement of other research by, 102, 125
  duplication of efforts, 13, 82, 105
  early estimates of costs for, 184-186
  economic impacts, 165, 172
  ethical considerations, 79-88
  expenditures, Federal, 8
  facilities, 10, 13
  focus, 7-9
  funding, *see* funding for genome projects
  interagency coordination and communications, 8, 11, 123
  interagency task force oversight of, 14, 119-121
  international efforts on, 133-159; *see also* specific countries
  lead agency concept, 12-14, 115, 116-118
  legislation, 12, 14, 123
  manpower availability, 10
  medical applications, 56-64; *see also* disease; medicine
  military applications, 174
  misconceptions about, 9-10
  national prestige associated with, 174
  objectives, 7, 9, 55
  organization of, 12-15
  organizations involved in, 6, 7
  policy development for, 134
  political interference with, 127-128
  quality control and reference standards, 103, 127, 183
  resource allocation for, 10; *see also* funding for genome projects
  scope of, 10, 134
  training of personnel, 183-184
  U.S. competitiveness and, 11, 133
  see also genome mapping; DNA sequences/sequencing; Human Genome Initiative; pilot projects
genome, human
  amount sequenced, 46-47

  bibliometric analysis of research on, 133, 157-158, 195
  size, 24, 43, 46-47
genomes
  bacteriophage T4, 40
  definition, 3, 21
  Epstein-Barr virus, 46
  mitochondrial, 71
  organization, 21
  regeneration, 21
  size, 21, 24-25, 43
  smallest, 148
genomic library, 35
Germany, *see* Federal Republic of Germany
Gilbert, Walter, 44-46, 126, 153, 156
glutamic acid, codon, 23
glutamine, codon, 23
glycine, codon, 23
granulocyte colony stimulating factor, 64
guanosine, 21-22
Gusella, James, 136

Harvard University, DNA sequencing, 44, 100
heart disease, 58, 62, 64
hemophilia, 56, 57, 58, 64
high-mobility group CoA reductase, 61
Hill, Lister, 97
histidine, codon, 23
Hitachi, Ltd., automated DNA sequencer, 47
Hood, Leroy, 47, 126
hormones, 21
Howard Hughes Medical Institute
  as lead agency for genome projects, 13
  budget, 8
  collaboration with CEPH, 146
  databases, 7, 8, 98, 106
  expenditures, 102, 106
  funding, 7, 105, 109
  genome initiatives, 8, 105-106, 109
  mission, 7
  RFLP mapping project, 6, 29
  university centers, 106
Hpa I, 28
Human Gene Mapping Library, 106, 189-190
Human Gene Mapping Workshop, 29, 106, 144, 157
Human Genetic Mutant Cell Repository, 31, 96, 190, 192-193
Human Genome Initiative
  budget, 7-8, 101
  expenditures, 7-8
  justification for, 102
  management, 6
  objectives, 6, 7, 14
  recommendations on, 101
  stages, 101
  workshops, 6
human growth hormone, 59, 62, 64
human physiology and development
  genome mapping applications to, 65
  NICHD-supported research, 95
Huntington's disease, 28, 57, 58, 64, 83, 134-136, 146, 149
hybridization, *see in situ* hybridization; somatic cell hybridization
hypercholesterolemia, 56, 58, 149
hypertension, 64

Imperial Cancer Research Fund, 148
in situ hybridization
  cDNA mapping by, 30, 33-34
  in mapping genes to whole chromosomes, 56
  localization of fruit fly clones by, 42, 43
Index Medicus, 97
Industrial Biotechnology Association, opinions on Federal initiatives in mapping and sequencing, 108
informatics
  Advanced Informatics in Medicine, 141
  Bioinformatics: Collaborative European Programs and Strategy, 141
  BIONET^{TM}, 193
  Contextual Measures for R&D in Biotechnology, 140-141
  National Biotechnology Information Center, 193
infrastructure for genome projects
  European, 141
  Federal support, 8, 102
  resource allocation, 10
Institute for Medical Research, somatic cell hybrid line repository, 31
insulin, 21, 59, 61, 62, 64
Integrated Genetics, DNA probe development, 58, 108
intellectual property, protection of, *see* patent and copyright policies
IntelliGenetics Corp., 96
interleukin-2, 62, 64
International efforts on genome projects
  collaboration and cooperation, 150-159; *see also* collaboration on genome research
  see also specific countries
International Geophysical Year, 150-151
isoleucine, codon, 23
Italy, human genome research, 8, 133, 145-147, 195

Japan
  automation of DNA sequencing equipment, 47-48, 137
  basic science expertise, 137
  collaboration on research, 157
  commercialization of mapping and sequencing technologies, 133, 138
  competitiveness with U.S., 133, 137-139
  cooperation with U.S., 139
  databases and repositories, 139
  expenditures on genome projects, 8-9, 138
  funding for genome research, 137
  grants program in genetics, 9
  Human Frontiers Science Program, 9, 137-138
  mapping and sequencing research, 136-138
  Ministry of Education, Science, and Culture, 9, 136-137
  Ministry of International Trade and Industry, 137-138

  peer review, 136
  physical mapping of *E. coli* genome, 41
  policy development on genome research, 136-137
  published genome research, 133, 195
  robotics technology, 137
  Science and Technology Agency, 8-9, 137
  workshop on DNA sequencing technologies, 47

karyotypes, human female, 34
karyotyping, 32-33, 56
Kennedy, John, 97
kidney diseases, 57, 58, 64
Kirschstein, Ruth, 94 n.1, 117
Koshland, Daniel, 126

Lalouel, Jean-Marc, 146
Latin America, genome research, 149
Lawrence Livermore National Laboratory
  chromosome sorting, 100
  mapping of chromosome 19, 44, 100
  ordering of DNA clones, 100, 108
Lederberg, Joshua, 126
leucine, codon, 23
leukemia, 64
Levinson, Rachel, 94 n.1
libraries
  of DNA fragments, construction of, 39
  of overlapping clones, 38-39
  see also repositories
life sciences
  DOE funding for, 7
  HHMI funding for, 7
  NIH funding for, 7
  NSF funding for, 7, 8
Lifecodes, DNA probe development, 58
ligase, 39
lily, haploid DNA content, 25
Lindberg, Donald A.B., 94 n.1
Los Alamos National Laboratory
  chromosome sorting at, 32, 100-101
  mapping of chromosome *16*, 44, 100
  ordering of DNA clones, 100
  see also GenBank®
Lovell, Joseph, 97
low-density lipoprotein receptor, 58, 61
lysine, codon, 23

macrophage colony stimulating factor, 64
Massachusetts Institute of Technology, bioprocess engineering center, 102
Maxam, Alan, 44-46
Max Planck Society, 144
McKusick, Victor, 24, 98
medicine
  diagnostic tool development, 56, 58-59
  drug development, 62-63
  human gene therapy prospects, 64
  isolation of genes associated with diseases, 59-62
  see also genetic diseases
meiosis, 26-27
Mendel, Gregor, 3, 73
Mendelian Inheritance in Man, 24, 98, 106, 189
Merriam, John, 42
messenger RNA
  function, 23
  size of human genes, 61
  translation into protein, 23, 30
methionine
  codon, 23
mitochondrial genome, human origin clues from, 71
molecular anthropology, 72
molecular biology
  Big Science vs. small science, 125, 126-127
  of human development, 65, 95
  manpower in, 10
  plant, genome mapping applications to, 73
molecular evolution
  genome mapping and DNA sequencing applications to, 57, 68-72
  human origins, 71
  primate, 70
  unanswered questions in, 69-70
monoamine oxidase, 86
Moskowitz, Jay, 94 n.1
Mount Sinai Medical Center Institute of Human Genomic Studies, 100
mouse
  beta globin gene, 65
  cell hybridization, *see* somatic cell hybridization
  genetic similarities to humans, 34, 67
  genetics database, 106
  haploid DNA content, 25
  Mus musculus, amount sequenced and genome size, 47
muscular dystrophy
  Becker's, 63
  Duchenne, 57-59, 61-63
mutations
  artificially induced, 25
  cancer from, 25
  chromosome structural changes involved in, 25-26
  deletion, 25, 31
  detection of, 42, 56, 58, 59
  duplication, 25
  in fruit flies, 42, 68
  Human Genetic Mutant Cell Repository, 31, 96, 190, 192-193
  human rates, 72
  inversion, 26
  lethal, 42
  in nucleotide sequence, 23
  saturating screen technique for, 42
  in sex cells, 25
  in somatic cells, 25, 31
  translocation, 25-26, 31

National Academy of Sciences
  recommendations on genome projects, 107
  role in genome project oversight, 14, 15, 124
  views on international cooperation, 153
  see also National Research Council

National Aeronautics and Space Administration, 150-151
National Biotechnology Information Center, 193
National Bureau of Standards, 103
National Cancer Institute, 93-94, 96, 98, 117
National Center for Biotechnology Information Act of 1986, 97
National Flow Cytometry Resource, 32, 97
National Institute of Allergy and Infectious Diseases, 93-94
National Institute of Child Health and Human Development, 93-94, 95,
National Institute of General Medical Sciences, 93, 95, 94
National Institute of Neurological and Communicative Disorders and Stroke, 93-94, 95
National Institute on Aging, 95
National Institute on Mental Health, 95
National Institutes of Health
  budgets for genome projects, 8, 93
  databases, *see* National Library of Medicine
  expenditures for genome projects, 8, 93, 109
  funding for genome projects, 6, 7, 14, 93-94, 95-97, 109, 155
  genome project objectives, 8, 93-95
  as lead agency for genome projects, 12, 13-14, 116-117
  mission, 7, 93
  organization, 93-94, 99, 116
  origin, 93
  peer review, 98-99, 109
  research infrastructure, 96-98
  research resources activities related to molecular genetics, 97
  see also specific institutes
  working group on human genome, 94
national laboratories, *see* specific laboratories
National Library of Medicine, 7, 8, 12, 97-98, 117
National Research Council
  physical mapping strategies, 44
  recommendations for genome projects, 4, 11, 107, 116
  sequencing strategies, 44
National Science Foundation
  as lead agency for genome projects, 14
  funding for genome-related research, 7, 8, 96, 102-103, 109
  mission, 7, 102
  role on genome projects, 116
nematodes
  Caenorhabditis elegans, 42, 47, 147-148
  genome mapping, 4, 42, 147-148
  haploid DNA content, 25
  human DNA sequences compared with, 68
  mutant, 67
neurofibromatosis, 149
nucleotides
  base order, 3, 21-22; *see also* DNA sequence/sequencing
  bonding between base pairs, 21
  dideoxynucleotide, 45
  mutations in sequence of, 23
  number of base pairs in human cells, 5
  substitution and recombination rates, 70
  total sequenced, 46
Office of Science and Technology Policy
  Biotechnology Science Coordination Committee, 104
  Committee on Life Sciences, 15, 104
  interagency coordination of genome projects under, 8, 104, 109, 124
  mission, 104
Office of Management and Budget, role in genome projects, 105
Office of Technology Assessment
  cost estimates for human genome projects, 180-185
  workshop on genome projects, 4
oncogenes, 67
Organization of American States, 149
Oudtshoorn skin disease, 134, 149

Palade, George, 94 n.1
pandas, 36, 69
parasites, 64
parathyroid, 61
Pardue, Mary Lou, 33
Pasteur Institute, 145, 157
patent and copyright policies, 16-17, 82, 88, 166-170
Patent and Trademark Amendments of 1980, 16, 187
peer review
  at DOE, 101, 109, 118
  by Japan, 136
  in a national consortium, 122
  at NIH, 98-99, 109, 118
Pepper, Claude, 97
Perkin-Elmer Cetus Instruments
  DNA sequencing technology, 45-46, 47
  probe development, 58
phenylalanine, codon, 23
phenylalanine hydroxylase, 61
phenylketonuria, 57
Philipson, Lennart, 142, 152
physical mapping/maps
  automation of, 47
  bacterial genome, 41
  bottom-up, 43
  chromosome sorting for, 31-32, 56
  comparative mapping of species, 34
  construction of libraries of DNA fragments for, 38-39
  contig, 30, 39, 40, 41, 43-44, 61
  costs, 181-182
  detail possible on, 37
  distance measurements, 27, 40
  forward genetics applications, 61
  fragmentation of DNA for, 37-39
  fruit fly, 42-43
  gene dosage technique, 34
  high-resolution, 30, 35-46, 56
  human genome, 35--41, 43-46
  in situ hybridization in, 33-34, 56
  karyotyping in, 32-33, 56
  linking of genetic maps with, 181-182
  low-resolution, 30-35, 41, 43, 56
  medical applications, 64
  nematode genome, 42
  NIH grants for, 84
  nonhuman genomes, 41--43

  ordering of clones for, 39, 40
  private enterprises, 108
  purification of chromosomal DNA for, 37
  rapid, mass-analysis approach, 41
  of restriction enzyme sites, 37-41
  size determinants, 43, 81
  somatic cell hybridization in, 27, 30-32, 33, 56
  strategies, 43-44
  time required for, 40-41
  top-down, 43
  yeast genome, 41-42
physicians' attitudes and practice, effects of genetic information on, 83
physiology, *see* human physiology and development
Pickett, Betty, 94 n.1
pilot programs
  DNA sequencing, 44
  European, 140, 145, 147
  importance, 4
  NRC recommendations for, 107
  yeast chromosome sequencing, 140
plants, genome mapping of, 73
polycystic kidney disease, 57, 58
polymerase, 45
population biology, genome mapping applications in, 72-73, 134
President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, 85
private sector
  foundation support, 109, 137, 147, 148
  funding of genome projects, 14-15, 108, 109
  genome mapping efforts, 9
  government formation of consortia with, 14-15
  role in genome projects, 107-108
  technology development, 108; *see also* automation; robotics
  see also specific companies
prokaryotes, lack of introns in, 69-70
proline, codon, 23
protein
  classification by period of invention, 68-69
  coding sequences, identification of, 65, 67
  databases, 97, 98, 158-159, 190-191
  engineering, 63, 142, 148
  folding, 66
  functions, 21-24, 67
  kinase C, 61
  life cycle regulator, 67
  for single-gene diseases, identification of, 57, 59
  structure/function relationship, 63, 65
  synthesis, 23-24, 30
  see also amino acids; enzymes; and other specific proteins
publications
  international, 157-158, 195

quagga, 72

recombinant DNA technology
  cloning vector development through, 36, 38
  definition, 24
  use to create genetic linkage maps, 3, 24, 28
repositories for research materials
  access to, 58, 134
  American Type Culture Collection, 101, 141, 190, 192
  Armed Forces Institute of Pathology, 104
  Center for the Study of Human Polymorphism, 33, 58
  clones, 97, 101, 115
  costs, 182
  DNA Segment Library, 97
  family pedigrees, 58-59
  Federal support, 8, 12, 96
  Human DNA Probes and Libraries, 96
  Human Genetic Mutant Cell Library, 31, 96, 190, 192-193
  importance, 4, 9
  international collaboration on, 158
  tissue, 104
reproductive choices, ethical considerations in, 83-84, 85
restriction enzyme cutting
  automation of, 47
  infrequent, 41
  partial, 39
restriction enzyme cutting sites
  on bacteriophage T4 genomic map, 40
  high-resolution physical mapping with, 35, 37-41
  polymorphisms, *see* restriction fragment length polymorphisms
restriction enzymes
  cDNA cutting with, 35
  Hpa I, 28
  Not I, 41
restriction fragment length polymorphisms (RFLP)
  allelic forms, 28
  diagnostic uses, 58, 59, 62
  DNA probe detection of, 28-29, 61-62
  linkage maps of, 28-30, 62
  mapping of, 28-29, 62, 73
  markers for genetic diseases, 28, 58, 62
retinoblastoma, 57, 59, 62
ribonucleic acid, *see* RNA
ribosomes
  functions, 23
RIKEN-Washington University collaboration, 157
RNA
  functions, 23-24
  ribosomal, 23, 30, 33
  see also messenger RNA; transfer RNA
  structure, 22
robotics
  devices for DNA sequencing, 48, 137
  DNA extraction devices, 108
  DOE initiative in, 7-8
  microchemical, 48, 137

Saccharomyces cerevisiae
  amount sequenced, 47
  genome mapping, 41-42
  genome size, 47
salamander, haploid DNA content, 24-25
Salmonella typhimurium, 45

Sanger, Fred, 44-45, 47
scanning tunneling microscopy, for DNA sequencing, 46
scleroderma, 64
Seiko, automation of DNA sequencing, 48, 138
Sendai virus, 31
SeQ, Ltd., physical mapping project, 108
serine, codon, 23
sex cells
  mutations, 25
  progenitors, 26
sickle cell disease, 4, 28, 56, 57, 58, 88
Singer, Maxine, 126
Sinsheimer, Robert, 100, 126
Smith, Cassandra, 41
Smith, Lloyd, 126
somatic cell hybrid lines, 31 n.2, 37
somatic cell hybridization
  cDNA mapping by, 30
  genome mapping applications, 27, 30-31, 33, 56
  of single copies of chromosomes, 31
somatic cells
  definition, 21
  mutations in, 25, 31 n.2
South Africa, genome research, 133, 149, 195
Soviet Union, genome research, 133, 149, 195
Sulston, John, 147-148
superoxide dismutase, 64

Task Force for Biotechnology Information, 141
Tay-Sachs disease, 4
technology development
  costs, 183
  NIH grants for, 94-95
  NRC recommendations for, 107
  private sector role in, 108
  see also automation; robotics
technology transfer
  advantages of national consortium for, 14-15, 121-122
  congressional encouragement of, 166
  economic implications, 165, 122, 172
  ethical issues, 87-88, 173-174
  international, 172-174
  military applications, 174
  national prestige issue, 174
  patent/copyright policies, 16-17, 82, 88, 122, 166-170
  strategies for improving, 16
  trade secrets, 170-172
Technology Transfer Act of 1986, 14, 16, 167
thalassemias, 57, 134
threonine, codon, 23
thymidine, 21-22
thyroglobulin, 61
Tinoco, Ignacio, 101
tissue plasminogen activator, 64
Trademark Clarification Act of 1984, 167
transcription
  of DNA into mRNA, 23-24, 25, 30
  of introns, 25, 30
transfer RNA
  functions, 23, 30
  structure, 23
translation of mRNA into protein, 23-24, 30
Trivelpiece, Alvin, 101
tryptophan, codon, 23
tumor necrosis factor, 62, 64
Turner's syndrome, 32, 64
tyrosine, codon, 23

United Kingdom
  DNA sequencing, 44
  equipment development, 48, 147-148
  expenditures, 147
  Medical Research Council, 42, 44, 147
  national genome research effort, 147-148
  nonhuman genome mapping, 8, 42, 147
  published genome research, 133, 195
University of California at Los Angeles, clone library, 42
University of California at Santa Cruz, workshop, 6, 100
University of Copenhagen Institute of Medical Genetics, 143
University of Helsinki, 144
University of Manchester Institute of Science and Technology, 48, 148
University of Wisconsin at Madison, physical mapping of *E. coli* genome, 41

valine, codon, 23
vectors
  cloning of, 36, 38-39
  cosmid, 36-37, 38, 42, 56, 101
  plasmid, 36-37, 38, 47, 56, 67, 101
  robotic devices for producing, 47
viral infections, 64

Wada, Akiyoshi, 152, 156
Walsh, James, 126
Washington University
  collaboration with RIKEN, 157
  physical mapping of yeast genome, 41
Watson, James D., 3, 21, 126, 152
Weinberg, Robert, 126
Wexler, Nancy, 135-136
White, Raymond, 29, 126, 146
Wilson, Allan, 126
workshops
  on automation of DNA sequencing, 47
  on collaboration for genome projects, 187
  on costs of genome projects, 188
  DOE-sponsored, 6, 97
  European Economic Community, 141
  International Human Gene Mapping, 29, 106, 144, 157
  information management system applications, 97
  on materials repositories and databases, 97
  Matrix of Biological Knowledge, 193-194
  NIH-supported, 96-97
  OTA, 187-188
  University of California at Santa Cruz, 6, 100
wound healing, 64
Wyngaarden, James, 93

yeast chromosomes
  artificial, 36, 39, 43, 56, 157
  electrophoretic separation of, 37
  genome mapping, 4, 41-42, 157
  haploid DNA content, 25
  lack of introns in, 69
  pilot project on sequencing, 140
  Saccharomyces cerevisiae, 41-42, 47
  similarities to other organisms, 67
  use to isolate human gene functions, 67, 68

Superintendent of Documents Publication Order Form

Order Processing Code:
***6394**

Charge your order.
It's easy!

☐ **YES,** please send me the following indicated publications:

Mapping Our Genes—The Genome Projects: How Big, How Fast?
GPO stock number 052-003-01106-9; price $10

1. The total cost of my order is $_______ (International customers please add an additional 25%.) All prices include regular domestic postage and handling and are good through 9/88. After this date, please call Order and Information Desk at 202-783-3238 to verify prices.

Please Type or Print

2. __________________________________
(Company or personal name)

____
(Additional address/attention line)

____
(Street address)

____
(City, State, ZIP Code)

(_______)__________________________
(Daytime phone including area code)

3. Please choose method of payment:

☐ Check payable to the Superintendent of Documents

☐ GPO Deposit Account ☐☐☐☐☐☐☐–☐

☐ VISA, CHOICE or MasterCard Account

☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐☐

____
(Credit card expiration date)

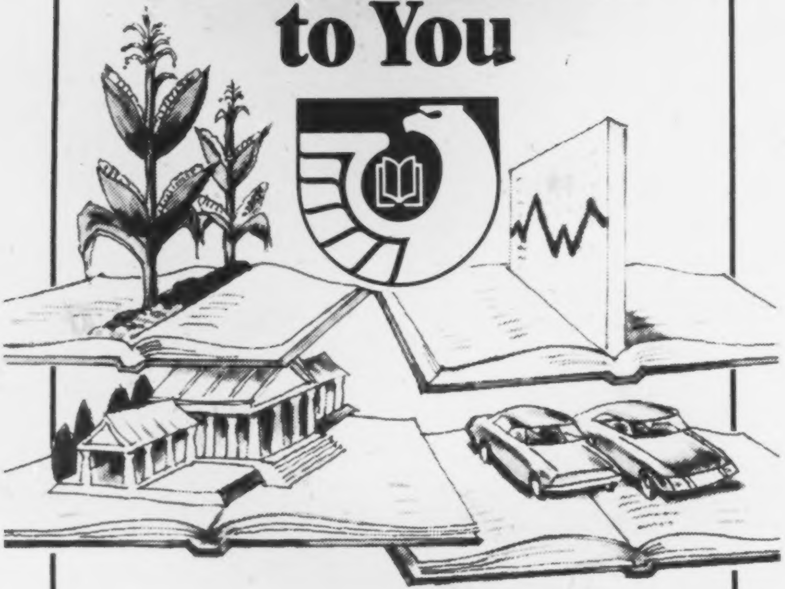
Thank you for your order!

____
(Signature)

4/88

4. Mail To: Superintendent of Documents, Government Printing Office, Washington, D.C. 20402-9325

Bringing Government Information to You



Information from the Federal Government, on subjects ranging from agriculture to zoology, is available at more than 1,380 Depository libraries throughout the United States.

These libraries allow you free access to thousands of publications issued by your Government and connect you to a variety of information resources to help answer your questions.

To locate the Depository Library in your area, contact your local library or write to the Federal Depository Library Program, Office of the Public Printer, Washington, DC 20401.

Federal Depository Library Program

This program is supported by The Advertising Council and is a public service of this publication.

U.S. GOVERNMENT PRINTING OFFICE : 1988 O - 83-360 QL 3

Office of Technology Assessment

The Office of Technology Assessment (OTA) was created in 1972 as an analytical arm of Congress. OTA's basic function is to help legislative policymakers anticipate and plan for the consequences of technological changes and to examine the many ways, expected and unexpected, in which technology affects people's lives. The assessment of technology calls for exploration of the physical, biological, economic, social, and political impacts that can result from applications of scientific knowledge. OTA provides Congress with independent and timely information about the potential effects—both beneficial and harmful—of technological applications.

Requests for studies are made by chairmen of standing committees of the House of Representatives or Senate; by the Technology Assessment Board, the governing body of OTA; or by the Director of OTA in consultation with the Board.

The Technology Assessment Board is composed of six members of the House, six members of the Senate, and the OTA Director, who is a nonvoting member.

OTA has studies under way in nine program areas: energy and materials; industry, technology, and employment; international security and commerce; biological applications; food and renewable resources; health; communication and information technologies; oceans and environment; and science, education, and transportation.

OTA-BA-373

APRIL 1988